Università degli studi di Napoli
Federico II

Tesi di dottorato in Statistica (XXVII Ciclo)

Tecniche di validazione per il Clustering di documenti

Maria Spano

Dipartimento di
Scienze Economiche e Statistiche

Aprile 2015

Tecniche di validazione
per il Clustering di documenti

Aprile 2015

III

Ringraziamenti

Ringrazio di cuore tutti coloro che mi sostenuto durante questo percorso di crescita professionale.

Maria

Indice

Introduzione

Classificare i documenti appartenenti ad un determinato *corpus*, sulla base del loro contenuto, è uno dei compiti più importanti, se non il più importante, del *Text Mining* e dell'*Information Retrieval*. Nel quadro delle tecniche per estrarre e organizzare la conoscenza contenuta in grandi basi documentarie, in maniera (semi-)automatica, al fine di soddisfare specifici fabbisogni informativi, risulta spesso determinante riuscire ad assegnare una medesima categoria, o etichetta, a documenti considerati "simili", dal punto di vista dell'argomento trattato.

L'assunto è che ogni documento presenta solo un sottoinsieme del vocabolario di termini del *corpus* e che documenti di categorie differenti presentano una diversa distribuzione della frequenza dei termini utilizzati. In forma più estrema, si assume che ciascuna categoria di documenti utilizza, in forma quasi esclusiva, una porzione del vocabolario, che costituisce un proprio (sotto)vocabolario peculiare, costituito da specifici termini. Ad esempio, all'interno di una raccolta di annate di un giornale, gli articoli che trattano di Finanza presentano tipicamente termini come *denaro*, *rendita*, *investimento*, mentre invece negli articoli che parlano di Sport si ritrovano termini come *gara*, *giocatore* o *arbitro*. Occorre sottolineare come i differenti vocabolari peculiari delle diverse categorie possano essere sovrapposti, così che ciascun

documento può utilizzare termini peculiari di categorie cui non appartiene. Questa circostanza rende il processo, noto come "categorizzazione" di un testo, estremamente delicato: non raramente documenti di fatto non appartenenti alla stessa categoria potrebbero mostrare una notevole similarità nel vocabolario utilizzato. Sempre in relazione ad articoli di giornale, si pensi, ad esempio, alla diffusione dell'utilizzo di termini sportivi nei discorsi dei politici (*scendere in campo*, la *squadra* di governo, il Presidente della Repubblica come *arbitro*).

L'obiettivo di raggruppare documenti simili in sottoinsiemi riconoscibili può essere raggiunto facendo riferimento a metodi di *Classificazione supervisionata* o *non supervisionata*. La scelta tra i due diversi approcci è funzione dell'obiettivo conoscitivo ed è vincolata dall'avere a disposizione, o meno, informazioni *a priori* da sfruttare nel corso del processo di classificazione.

Le informazioni *a priori* riguardano, in genere, il numero di gruppi esistenti nel *corpus* analizzato, le loro peculiarità e la loro composizione. Nel caso di un approccio di *Classificazione supervisionata* queste informazioni si basano di solito sulla conoscenza di esperti che consente, prima del processo di classificazione, l'assegnazione di una etichetta (categoria) agli elementi di un sottoinsieme di documenti, in modo da definire una procedura per attribuire automaticamente la stessa categoria a documenti simili. Al contrario un approccio *non supervisionato* mira a raggruppare i documenti cercando di far emergere la naturale struttura delle categorie, non nota *a priori*. Il processo di classificazione, detto, in questo caso, *Clustering*, si basa unicamente sui i dati a disposizione.

Si noti che la nozione di metodi di *Classificazione supervisionata* e *non supervisionata* non è certo specifica di problemi in cui gli elementi da raggruppare siano documenti, ma questo è il contesto in cui si pone tale lavoro,

introducendo le implicazioni connesse alla particolare natura dei dati testuali.

In letteratura, esistono numerosi studi comparativi che mostrano una migliore *performance* dell'una o dell'altra famiglia di metodi, ignorando, talvolta, il diverso obiettivo che essi si pongono e il diverso contesto in cui operano: un conto è disporre di informazioni relative all'esistenza ed alle caratteristiche delle diverse categorie di documenti, altro conto è, come accade in numerosi ambiti applicativi, dover procedere sulla base della sola informazione contenuta negli stessi documenti da categorizzare.
É la stessa natura ambigua, e spesso polivalente, dell'informazione testuale a fare frequentemente preferire un approccio *data-driven*, "non supervisionato", anche in presenza di categorie costruite sulla base di valutazioni esterne. Quando infatti le categorie sono scelte e attribuite sulla base di conoscenza esperta, non c'è sempre (o c'è solo parzialmente) una valutazione del contenuto lessicale complessivo dei documenti e, ancor meno, del contenuto semantico (basti pensare ad esempio all'utilizzo di metafore o di altre figure retoriche in un contesto non prettamente letterario).
Nelle tecniche supervisionate si ha di fatto una replicazione del lavoro di operatori umani esperti, con l'implementazione di una fase di apprendimento per l'assegnazione degli oggetti da classificare a categorie definite. Il rischio in cui si può incorrere è forzare una classificazione che in realtà non è naturalmente presente nello specifico contesto analizzato. Ma il medesimo rischio si può correre ovviamente anche per il *Clustering*, poiché, comunque, nella ricerca della miglior partizione di un insieme di oggetti – quindi di documenti, nel caso di una base di dati documentaria – si presuppone che esista una naturale classificazione, non nota al ricercatore, da fare "emerge-

re" grazie ad un'analisi di tipo statistico.

Nell'ambito delle tecniche di *Clustering* sono stati proposti nel tempo una grande quantità di soluzioni. Le ragioni del grande interesse della comunità scientifica sono sostanzialmente due: da un lato, la polivalenza di questo approccio e la sua vasta applicabilità ha portato allo sviluppo di numerosi metodi applicabili a dati dalle specificità più diverse (misure, attributi, testi, immagini), dall'altro, l'evoluzione tecnologica ha indirizzato la ricerca nella direzione di un affinamento delle componenti algoritmiche e computazionali, ancor più che di quelle specificamente metodologiche.
Pertanto, da ciascuno dei metodi di *Clustering*, di regola caratterizzati dall'ottimizzazione del criterio in base al quale sono individuati i raggruppamenti, sono derivati una serie di algoritmi, realizzati per rendere il processo di classificazione più efficiente in termini di calcolo e per essere in grado di trattare dati altamente dimensionali.
Lo stato dell'arte in tema di Classificazione automatica è fornito da una letteratura in continua e velocissima crescita, poco integrata, poiché sviluppata in contesti differenti e con differenti linguaggi e notazioni, all'interno della quale un ricercatore, non necessariamente poco esperto, ha molte difficoltà nella scelta del metodo di *Clustering* più adatto al suo problema specifico – ammesso che questo metodo "adatto" esista e sia unico. Inoltre, una volta che abbia, comunque, scelto il metodo, non ha a disposizione uno strumento che gli consenta di valutare oggettivamente la qualità della soluzione ottenuta. Come non esiste un algoritmo di *Clustering* "ottimo", in quanto i diversi metodi generano in qualsiasi caso una suddivisione, anche quando i dati non presentano alcun raggruppamento naturale, così non esiste uno strumento che, univocamente, sappia valutare il successo delle

applicazioni di *Cluster Analysis*, ovvero che consente di sapere se il modello di raggruppamento individuato corrisponde ad una struttura reale o meno. Un rischio ulteriore è rappresentato dalla circostanza che l'insieme delle tecniche che vanno sotto il nome di *Cluster Analysis*, è connotato come parte del più ampio contenitore dei metodi di tipo esplorativo, in quanto finalizzati a identificare strutture nei dati non note. La validazione dei risultati viene, quindi, a volte, percepita come un'aggiunta onerosa a quello che viene considerato un passo iniziale di un processo di conoscenza. Questa percezione è estremamente pericolosa, perché potrebbe indirizzare gli approfondimenti successivi in maniera non corretta.
La motivazione del presente lavoro è, quindi, data dalla consapevolezza dell'importanza di disporre di strumenti di valutazione efficaci, così da fornire al ricercatore risultati caratterizzati da un certo grado di affidabilità. Il problema della validazione della struttura di partizionamento ottenuta e dell'effettiva interpretabilità dei gruppi risultanti è, quindi, centrale ed è l'oggetto del presente lavoro.
Questa tesi pone, quindi, in rassegna la vastissima letteratura relativa agli strumenti di validazione, partendo dall'assunto che essi debbano essere indipendenti dall'algoritmo scelto per il raggruppamento, debbano essere utili per individuare il numero di gruppi presenti nei dati, per valutare se i gruppi ottenuti sono significativi (o sono solo un artefatto degli algoritmi), o ancora per decidere quale tra i diversi algoritmi scegliere.
L'insieme delle tecniche rivolte alla valutazione quantitativa e oggettiva dei risultati di una classificazione sono note come *Cluster validation methods*.
É possibile considerare due approcci a seconda che si faccia o meno riferimento ad informazione *a priori* nel processo di validazione:

1. I criteri di validazione esterni valutano la soluzione ottenuta compa-

randola con l'informazione *a priori* riguardante le vere etichette di classe. Infatti, molti di questi indicatori sono di fatto utilizzati nell'ambito della classificazione supervisionata, e riadattati a problemi di *Clustering*. È chiaro che un approccio del genere ha senso soprattutto quando si vuole testare un algoritmo o comunque in un esperimento controllato, poiché nelle applicazioni su dati reali la struttura non è nota e di conseguenza non è disponibile la loro partizione corretta.

2. Quando l'appartenenza degli oggetti a categorie pre-specificate non è disponibile la validazione della soluzione di *Clustering* ottenuta può avvenire seguendo diverse strade:

   - cercando di valutare a monte se i dati siano effettivamente raggruppabili (*Cluster tendency*);
   - valutando la stabilità dell'algoritmo rispetto a differenti campioni della base di dati di partenza, con il vantaggio di essere indipendenti dal metodo di *Clustering* scelto, pur non consentendo però una validazione diretta della soluzione ottenuta (*Cluster stability*);
   - facendo riferimento a misure che valutino la compattezza e la separazione dei gruppi individuati (*Cluster validation*).

In un processo di *Clustering*, in cui non è nota *a priori* alcuna informazione circa la composizione dei gruppi, i criteri interni per la validazione rappresentano l'unica opzione disponibile per valutare la qualità della soluzione. Date le differenze tra i tre approcci sopra menzionati non risulta possibile compararli riferendoci ad uno stesso quadro di riferimento. In questo lavoro l'attenzione è stata focalizzata sulle *misure interne per la validazione*,

poiché rappresentano degli strumenti in grado di fornire un'informazione sintetica della qualità della soluzione ottenuta rispetto ad altri approcci che seppur ugualmente validi comportano una serie di operazioni non banali per ottenere un risultato definitivo e che oltretutto non consentono di valutare in modo diretto tale qualità.
Così come la varietà dei metodi di *Clustering* discende principalmente dalla loro applicabilità in molti ambiti diversi, così diverse misure interne di validazione sono state prodotte in diversi contesti. Questa circostanza rende necessario per i ricercatori avere a disposizione delle linee guida per districarsi al meglio nelle scelte di uno strumento piuttosto che un altro che possa risultare più idoneo nei diversi contesti.
In questo lavoro si propone un confronto dei punti di forza e dei punti di debolezza di numerosi indici, tra i più utilizzati e più recenti, valutandone la *performance* su un gran numero di configurazioni. Ne deriva la consapevolezza che la proposta di un nuovo indice sarebbe una operazione di poca efficacia. Si preferisce piuttosto perseguire l'obiettivo di individuare una strategia integrata che metta in relazione i tre elementi dai quali non si può prescindere per eseguire una *Cluster Analysis*: il tipo di dati, gli algoritmi utilizzati e gli indici di validazione. Questi elementi condizionano in maniera determinante vuoi la qualità dei risultati dell'analisi, vuoi la valutazione che se ne può ottenere. L'ambito applicativo prescelto è quello del *Clustering* di Documenti, con le peculiarità e le specificità già descritte.

Struttura della tesi

Nel **primo capitolo** si presenta una descrizione di tutti gli aspetti connessi alla definizione e all'organizzazione dei dati testuali, con particolare atten-

zione alle diverse procedure di trasformazione del testo in un insieme di dati strutturati, approfondendo i differenti step attraverso i quali un documento scritto in linguaggio naturale possa essere codificato per poter essere trattato con tecniche statistiche.
Si presenta inoltre una panoramica dei diversi approcci per ridurre la dimensionalità del vocabolario, in termini di tecniche di *feature selection* e di *feature transformation*. Le prime rappresentano degli strumenti utili al fine di selezionare i termini che maggiormente rappresentano il contenuto dei documenti, e che quindi maggiormente li caratterizzano e li discriminano al meglio.
L'alternativa a questo approccio è ridurre l'elevata dimensionalità dello spazio definito dai termini del vocabolario attraverso l'impiego di metodi di *feature transformation*. L'idea di fondo di questo tipo di metodi è operare una riduzione dimensionale definendo nuove caratteristiche che siano una rappresentazione funzionale delle caratteristiche dell'insieme dei dati d'origine.

Nel **secondo capitolo** è estensivamente esposta una rassegna sulla *Cluster Analysis*, descrivendo le diverse fasi che costituiscono un processo di *Clustering*, e ponendo l'attenzione sulle diverse scelte che un ricercatore deve operare lungo il percorso e come queste influenzino i risultati ottenuti.
In particolare sono presentate le peculiarità specifiche del *Clustering* di Documenti, con le diverse misure di similarità e di distanza utilizzabili in questo contesto, e i principali criteri dai quali sono stati sviluppati la maggior parte degli algoritmi considerati in letteratura più efficaci per questo tipo particolare di dati.

Il **terzo capitolo** riguarda le tecniche di validazione nel *Clustering*. Nello specifico sono presentati i diversi approcci per la validazione di una soluzione di *Clustering*, in termini di misure interne ed esterne per la validazione, che consentono di confrontare tra differenti schemi e valutare la qualità di una soluzione con e senza avere a disposizione informazioni *a priori*. Successivamente viene descritto il modo in cui le misure interne di validazione sono utilizzate per la ricerca della partizione ottimale, in termini di numero di classi presenti nel dataset di partenza.
Infine sono presentati per completezza due ulteriori approcci alla validazione, noti come valutazione della *Cluster tendency* e della *Cluster stability*. Il primo mira a valutare la tendenza dei dati ad essere raggruppati, attraverso l'uso di test statistici rivolti all'individuazione di una struttura non casuale nei dati.
La *Cluster stability* mira invece a stabilire quanto è robusta una soluzione sotto perturbazione o sub-campionamento dei dati originali. Una partizione o una gerarchia è considerata stabile quando "cattura" la struttura sottostante un insieme di dati, sotto l'assunzione che tale struttura possa essere riproposta in altri dataset, tratti dalla stessa origine.

Nel **quarto capitolo** sono presentati infine i risultati di uno studio comparativo sugli indici di validazione in modo da descriverne criticamente il funzionamento e le specificità. Sono proposti i risultati ottenuti con differenti algoritmi di *Clustering* sia su dati simulati sia su dati reali di tipo quantitativo. A partire da tale studio sono poi state analizzate le performance degli indici con gli algoritmi più idonei per il *Clustering* di documenti, in modo da poter approntare un secondo studio comparativo su basi di dati di tipo testuale. L'analisi dei risultati ottenuti, alla luce anche della più

diffusa letteratura, fornisce interessanti spunti critici rispetto al problema trattato.

Capitolo 1

Linguaggio naturale e dati testuali

Il *linguaggio naturale* è il mezzo attraverso il quale l'uomo esprime i propri pensieri, le proprie opinioni, ed interagisce con il mondo esterno. Esistono differenti mezzi per esercitare questa capacità, ma dal momento in cui gli esseri umani hanno iniziato a trasporre il linguaggio anche in forma scritta, il *testo* è diventato il modo più comune e diffuso per memorizzare, comunicare e scambiare informazioni.
Lo studio del linguaggio naturale non ha certo origine recente. Secondo K. Krippendorff [49] la traccia più antica di cui si ha notizia risale al Diciottesimo secolo, ed è relativa all'analisi di una raccolta di novanta inni intitolata *Canti di Sion*, realizzata in Svezia da un autore sconosciuto.
Fino alla prima metà del XX secolo lo studio del linguaggio naturale, inteso come capacità di espressione dell'uomo a un determinato livello comunicativo, era tradizionalmente campo di ricerca da parte di linguisti, psicologi e sociologi. I primi studi quantitativi sono stati quindi sviluppati con l'o-

biettivo specifico di codificare le regolarità esistenti nella lingua [92] [90]. A partire dagli Sessanta lo studio di dati espressi in linguaggio naturale è stato via via oggetto d'interesse di molteplici altri ambiti disciplinari, tra cui la Statistica. In particolare, ai rappresentanti della scuola francese di Analisi dei Dati va l'indubbio merito di aver introdotto numerose proposte metodologiche e computazionali di particolare interesse. Come lo stesso J.P. Benzécri precisa in *Histoire et Préhistoire de l'Analyse des Données*, ad esempio, la stessa Analisi delle Corrispondenze "è stata inizialmente proposta come un metodo induttivo per l'analisi di dati linguistici" [7].
Intorno alla fine degli anni '80, le tecniche statistiche per l'analisi di dati espressi in linguaggio naturale hanno subito forti cambiamenti, strettamente legati all'evoluzione dell'Informatica, portando tra l'altro allo sviluppo dell'Analisi automatica dei testi e della Statistica testuale [54] [55]. La crescente e continua produzione e diffusione di dati testuali in formato digitale, oltre che la necessità di metodologie per l'acquisizione, classificazione e gestione automatica di grandi basi di dati, ha poi dato vita al più recente e noto dominio di ricerca del *Text Mining*.
Per sua natura il linguaggio è un fenomeno complesso, in continua evoluzione e difficile da analizzare con procedure automatiche. Il *Text Mining* è quindi necessariamente un campo altamente interdisciplinare, il cui progresso è legato al contributo congiunto di diversi settori scientifici, quali la Linguistica, l'Informatica e la Statistica.
Il *Text Mining* si propone sviluppare procedure per estrarre conoscenza e soddisfare i diversi bisogni informativi degli utenti. Nello specifico un processo di *Text Mining* può essere rivolto a: visualizzare l'informazione nascosta nei testi, reperire informazioni rilevanti data una query definita dall'utente, al *browsing* di documenti, o ancora a classificarli in categorie

con o senza informazione *a priori*. Per ciascuno di questi compiti sono state sviluppate una serie di metodologie.

1.1 Strutturare il non strutturato

Il testo è un oggetto in qualche misura sfuggente. Ogni testo, pur nascendo da strutture e da un universo culturale condiviso, ha una originalità propria, qualcosa che lo rende unico. Analizzare un testo da un punto di vista quantitativo può essere quindi molto difficile a causa della complessità e vastità del linguaggio naturale e delle ambiguità in esso presenti.
La difficoltà intrinseca del trattamento dei testi deriva dalla necessità di dare una struttura e organizzazione ai dati prima di procedere con una qualunque analisi quantitativa.
Questa trasformazione del testo consiste nella selezione dei termini che riescono a rappresentare l'asse sintagmatico del documento. In letteratura questa fase è nota come *pre-trattamento del testo*, ma non esiste una definizione univoca delle operazioni in esso incluse. Quindi è il ricercatore che, in conformità con i propri obiettivi di ricerca, crea una strategia *ad hoc* per la comprensione delle informazioni contenute nella raccolta di testi che sta analizzando. E' facile capire come le scelte del ricercatore possano condizionare notevolmente i risultati dell'analisi stessa.
Una volta definiti gli obiettivi dell'analisi, così come si procede per una classica indagine statistica, deve essere innanzi tutto identificato il *collettivo* oggetto di studio. In tal senso va selezionata una raccolta coerente di testi, detta *corpus*, omogenea sotto un qualche punto di vista. Un *corpus* può essere quindi composto da documenti, interviste, rassegne stampa,

risposte a domande aperte in questionari, forum, blog, organizzati in una collezione.
Se in un'indagine tradizionale le caratteristiche osservate sono rappresentate tanto da informazioni riguardanti gli aspetti socio-demografici degli individui quanto da informazioni specifiche, rilevate per ottenere un quadro complessivo del fenomeno indagato, quali sono le variabili di un *corpus*? Perseguendo l'obiettivo di evidenziare in maniera automatica le tematiche e i significati prevalenti, è chiaro come le variabili oggetto di studio siano proprio i termini in esso utilizzati. In generale un termine (o parola) può essere definito come una sequenza di caratteri cui è associabile un significato. Questa definizione è in realtà molto vaga, mentre per condurre un'analisi statistica in senso stretto è necessario definire univocamente quale sia l'unità d'analisi da prendere in esame.
Prima ancora di fare ulteriori considerazioni su quale "tipo" di termine (unità di analisi) sia più idoneo per l'obiettivo di ricerca prefissato, si procede alla costruzione del *vocabolario* del *corpus*. Questa operazione viene effettuata trasformando il testo in una lista di termini, così come si presentano nella raccolta in esame, cui viene associato il valore della rispettiva occorrenza (numero di volte in cui un termine si presenta nella raccolta). Il *vocabolario* rappresenta in qualche modo la *distribuzione statistica* dei termini all'interno della collezione.
L'ampiezza (o dimensione) del vocabolario V è definita dal numero di termini presenti nella raccolta analizzata:

$$V = V_1 + \ldots + V_k + \ldots + V_{f_{max}} \tag{1.1}$$

dove V_1 rappresenta il numero di termini che si presentano una volta sola all'interno del testo (i cosiddetti *hapax*), V_k il numero di termini che si

presentano k volte, e $V_{f_{max}}$ il numero di termini con il maggior numero di occorrenze nel vocabolario.
É facile immaginare come anche per una raccolta di dimensioni contenute il vocabolario possa raggiungere un'ampiezza non sempre facilmente gestibile. L'obiettivo delle procedure di pre-trattamento, dalla definizione dell'unità di analisi alla costruzione della tabella lessicale, è pertanto non solo trasformare il *corpus* in una forma strutturata trattabile con tecniche statistiche, ma anche ridurre la variabilità, risolvendo le ambiguità presenti nel testo.

1.1.1 La definizione dell'unità di analisi

Gli studi quantitativi sul linguaggio nel corso del tempo hanno generato un vivace dibattito già solo sulla definizione dell'unità d'analisi.
Gli statistici di tradizione "formalista" [6][70][67] mostrano che partendo da un'analisi puramente formale si arriva a cogliere la struttura del senso presente nel *corpus*, e pertanto suggeriscono di considerare come unità di analisi la *forma grafica*. Con tale termine s'intende una sequenza di caratteri delimitata da due separatori. Secondo questa definizione si connota un qualsiasi termine nel testo secondo la sua grafia originale, indipendentemente dal significato e soprattutto dalla lingua. Tuttavia, spesso una forma grafica può risultare ambigua finché non si estende il contesto: ad esempio -abito- può riferirsi sia a -io abito- [verbo], sia a -l'abito- [sostantivo].
In contrapposizione ai formalisti, i linguisti quantitativi di tradizione harrisiana [33] sviluppano strumenti concreti di linguistica computazionale come i dizionari elettronici e i lessici di frequenza [19], definendo come unità di analisi il *lemma*. Il lemma è la *forma canonica* corrispondente all'entrata del termine nel dizionario e rappresenta tutte le flessioni con cui quell'unità lessicale può presentarsi nel discorso. Ad esempio –andare- è il lemma di

varie forme grafiche flesse, quali -andavamo-, -andiamo-, -vai- e così via. Sebbene la scelta del lemma come unità d'analisi può risolvere l'ambiguità di alcune forme grafiche (nell'esempio precedente la forma grafica -abito- se nel testo è un verbo diventa -abitare- se sostantivo resta -abito-), possono nascere altre ambiguità: la forma canonica -essere- da sola può riferirsi a due diversi lemmi, -essere- [verbo] e -essere- [sostantivo].
Nell'ambito della Statistica Testuale un buon compromesso tra le due visioni è considerare un'unità di analisi di tipo misto, che S. Bolasco definisce *forma testuale* [11]. Una forma testuale è una componente significativa minima del discorso non ulteriormente decomponibile. In quest'ottica l'unità di analisi non è più indipendente dalla lingua, ma può trattarsi di una forma grafica, di un lemma o di un poliforme, cioè tutte quelle sequenze di termini che esprimono un contenuto autonomo, quali ad esempio -Capo dello Stato-, -Presidente del Consiglio-.

Le diverse fasi di preparazione del *corpus* non costituiscono congiuntamente una strategia da seguire pedissequamente, dato lo stretto legame con la scelta dell'unità di analisi. Infatti, a seconda dell'obiettivo, il ricercatore può concepire una strategia *ad hoc* per il trattamento della raccolta in esame. La possibilità di considerare un'unità d'analisi di tipo misto, come la *forma testuale*, richiede, infatti, di pretrattare il *corpus* seguendo una serie di regole proprie del linguaggio naturale, che si concretizzano in una serie fasi operative:

Il **parsing** consente di individuare le successioni di caratteri dell'alfabeto compresi tra due separatori, in modo da ottenere una lista di forme grafiche. Con questa operazione il testo viene quindi frammentato, perdendo tutto ciò che concerne il contesto, lo stile e soprat-

tutto il significato di quest'ultimo, sacrificando pertanto una parte d'informazione.

La **normalizzazione** agisce sull'insieme dei caratteri non separatori, eliminando le possibili "replicazioni" del dato, come ad esempio le forme grafiche con lettera iniziale maiuscola o minuscola. Inoltre, attraverso questa procedura è possibile uniformare le forme che presentano forte variabilità, come ad esempio date, sigle e nomi propri.

L'**estrazione dei segmenti (lessicalizzazione)** consiste nell'individuare nel testo, fissando *a priori* una soglia di frequenza, i cosiddetti segmenti ripetuti. Questi ultimi sono disposizioni di $2, 3, \ldots p$ forme grafiche che si ripetono più volte all'interno del *corpus*. Tali sequenze possono essere vuote o incomplete, ossia formate solo da parole grammaticali, oppure caratteristiche, nel caso in cui costituiscono unità di senso indipendenti. In quest'ultimo caso si parla di poliformi, che vanno lessicalizzati per poter essere trattati come un'unica unità lessicale.

Il **tagging grammaticale** è una fase di annotazione del testo che consiste nell'attribuire a ciascun termine la categoria grammaticale di appartenenza (sostantivo, articolo, verbo, avverbio, aggettivo, pronome, preposizione). Con questo tipo di operazione è possibile disambiguare le forme polisemiche o omonime. La polisemia è frutto dello sviluppo nel tempo di una cultura, in quanto per esprimere nuovi concetti anziché coniare nuove parole. Può accadere infatti che vengano attribuiti nuovi significati ad un significante preesistente, come ad esempio la parola -farfalla- che può indicare un insetto o un componente meccanico. Per omonimia, o più specificatamente omografia, si intende una stessa forma grafica che può essere ricondotta a più lemmi, ad esempio

la parola -fine- può indicare un sostantivo maschile, femminile o un aggettivo.

Il **tagging semantico** consiste nell'etichettare i termini del vocabolario con meta-informazioni di tipo semantico. Un esempio di tagging semantico potrebbe essere l'attribuire, in uno studio sull'immigrazione, l'etichetta "nazionalità" alle forme -italian*-, -marocchin*-, -rumen*-.

La **lemmatizzazione** consiste nel riportare ogni forma flessa al lemma di appartenenza. I lemmatizzatori automatici, sono strumenti che possono raggiungere elevatissimi livelli di qualità nell'individuazione del giusto lemma, identificando nei testi strutture e regole capaci di definire univocamente diverse funzioni grammaticali.

É bene precisare che nell'ambito del *Text Mining*, le difficoltà dovute all'enorme peso computazionale "costringono" ad operare per lo più sulle forme grafiche (termini, parole), o piuttosto che a lemmatizzare automaticamente il testo, a fare riferimento alle radici dei termini attraverso il cosiddetto *stemming*. Questa procedura rispetto alla lemmatizzazione ha il vantaggio di essere computazionalmente più efficiente, in quanto non richiede l'utilizzo di risorse esterne, tuttavia con un pretrattamento del genere si corre comunque il rischio di considerare come unica occorrenza termini con una radice comune ma di significato sostanzialmente diverso. In definitiva, rispetto al tipo di unità di analisi scelta, tutte o solo alcune delle fasi appena descritte saranno poste effettivamente in essere per pretrattate la raccolta di documenti presa in esame [12].

1.1.2 La codifica

Dopo aver identificato l'unità di analisi coerentemente con gli obiettivi della ricerca, è necessario organizzare i documenti in una rappresentazione strutturata. I documenti necessitano, quindi, di essere sottoposti ad una fase di codifica, associando ciascuno di essi ad una rappresentazione compatta dei suo contenuto. Uno dei più popolari schemi per codificare *corpora* testuali in linguaggio naturale è il *Bag-of-Words* (BOW). Già dal nome del sistema di codifica si deduce che ogni documento verrà rappresentato da una collezione non ordinata di termini, di cui si trascura la grammatica e l'ordine dei termini stessi. Attraverso questa trasformazione ciascun documento D_j della raccolta è rappresentato da un vettore nello spazio definito dai termini del vocabolario:

$$D_j = (w_{1,j}, w_{2,j}, w_{3,j}, \ldots, w_{p,j}) \tag{1.2}$$

dove ogni elemento del vettore $w_{i,j}$ è il peso della *i-mo* termine nel *j-mo* documento, con $i = 1 \ldots p$ e $j = 1 \ldots n$.
Il principale vantaggio della codifica BOW è la relativamente bassa complessità computazionale, derivante dal fatto di poter trattare i documenti analizzati come vettori e quindi calcolare facilmente delle misure statistiche d'interesse. Seguendo lo schema *Bag-of-Words*, i vettori/documento possono essere poi giustapposti per costruire una matrice T, detta *tabella lessicale*, con n righe e p colonne. Sulle righe sono rappresentati gli n documenti della raccolta, mentre sulle colonne si trovano i p termini del vocabolario del *corpus*, individuati dopo aver effettuato le operazioni di pretrattamento. In accordo con uno schema di ponderazione, nella tabella (*documenti* $\times$ *termini*) (come nell'esempio riportato nella figura 1.1) ogni cella f_{d_j,t_i} potrà contenere tanto la frequenza dell'*i-mo* termine nel *j-mo* documento quanto

semplicemente la presenza/assenza, o ancora una funzione della frequenza stessa, come specificato nel paragrafo successivo.

	termine 1	termine 2	termine 3	$\cdots$	termine p
doc 1	1	0	0	$\cdots$	1
doc 2	0	2	1	$\cdots$	0
doc 3	1	0	1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
doc n	1	0	0	$\cdots$	0

Figura 1.1: Tabella lessicale

Facendo riferimento alla terminologia statistica, la matrice così costruita può essere vista come una tabella di contingenza, pur considerando i documenti come unità/osservazioni e le parole come modalità della variabile lingua. In questa forma strutturata diventa comunque possibile analizzare una raccolta di documenti con tecniche analoghe a quelle proposte per i classici dati numerici.

1.1.3 Scelta dei pesi

I termini hanno una diversa importanza nel descrivere il contenuto semantico dei documenti. Spesso, le informazioni più importanti e ricche di contenuto sono rappresentate da termini che si presentano nel documento raramente. Nella maggior parte dei casi, i termini che si presentano più frequentemente nel documento non sono invece interessanti, poiché ad esempio servono per collegare diverse parti del discorso o per esprimere concetti generali che non hanno legame con il contenuto vero e proprio del documento stesso.

Pertanto, contestualmente alla fase di codifica dei documenti si dovranno scegliere dei pesi da attribuire a ciascuno dei termini selezionati nella fase di pretrattamento. La scelta del sistema di ponderazione da utilizzare è, quindi, un aspetto fondamentale nel trattamento statistico dei dati testuali, in quanto i pesi dovranno riflettere l'importanza, in apporto informativo, di ogni termine presente nei documenti.
I principali schemi di ponderazione per quantificare numericamente l'importanza di un termine sono i seguenti:

- **Booleano**, dove il peso $w_{i,j}$ assume valore 1 se il termine i è presente nel documento j, altrimenti assume valore 0;
- **Frequentista**, dove il peso $w_{i,j}$ è pari ad $n_{i,j}$, cioè la frequenza del termine i nel documento j;
- **Frequentista normalizzato**, dove il peso $w_{i,j}$ è pari a $n_{i,j}/max\ n_j$, con $max\ n_j$ frequenza della parola che si presenta più volte nel documento j;
- **Term Frequency/Inverse Document Frequency** (TF/IDF), schema proposto originariamente da Salton e Buckley nel 1988 [71], e spesso utilizzato nell'ambito dell'*Information Retrieval*.

Lo schema di ponderazione *booleano* è il più semplice, poiché in tal caso si valuta solo presenza o assenza di un determinato termine nel singolo documento. Il limite evidente di tale sistema di ponderazione è che l'importanza di ogni termine non è ben misurata, in quanto è espressa in egual modo tanto nei documenti fortemente caratterizzati da esso quanto nei documenti in cui lo stesso termine non ha un contenuto informativo caratterizzante.

In genere nell'analisi di *corpora* testuali si preferisce uno schema di ponderazione *frequentista*, che tiene conto del numero di occorrenze di ogni termine all'interno del documento. Tale quantità è in realtà assimilabile ad una frequenza assoluta, poiché è calcolata considerando il numero di volte in cui il termine i occorre in ogni documento.
L'informazione statistica contenuta nelle distribuzioni marginali di colonna di una tabella lessicale ($documenti \times termini$) assume significato diverso, a seconda che si scelga lo schema di ponderazione *booleano* o quello *frequentista*. Nel primo caso infatti rappresenta il numero di documenti nel quale è presente ogni dato termine, mentre nel caso di uno schema *frequentista* rappresenta il numero di volte in cui ogni termine è presente all'interno della raccolta analizzata, quindi è di fatto associabile al vocabolario.
Il tipo di codifica prescelto dipende dal tipo di analisi che si vuole effettuare sui dati. In alcune strategie di trattamento del linguaggio naturale si preferisce utilizzare dei sistemi di pesi complessi. Come detto, nell'ambito dell'*Information Retrieval* G. Salton e C. Buckley propongono il *Term Frequency/Inverse Document Frequency*, schema di ponderazione che tiene conto al tempo stesso dell'importanza di ogni termine rispetto ad uno specifico documento e rispetto a tutti i documenti contenuti nel *corpus*.
L'idea di fondo alla base di questo schema scaturisce da alcune considerazioni sul trattamento dell'informazione testuale. I termini più frequenti all'interno di un documento sono considerati generalmente indicativi del contenuto del documento stesso, fatta eccezione per le parti del discorso funzionali (si veda il paragrafo 1.1.4). Per tener conto dell'importanza relativa di ogni termine è opportuno utilizzare, come fattore di normalizzazione, il numero di occorrenze del termine che appare più volte all'interno del

documento. Tale rapporto rappresenta il cosiddetto *term frequency*:

$$tf_{ij} = 0.5 + 0.5\frac{f_{ij}}{\max f_j} \tag{1.3}$$

Considerando un campo di variazione compreso tra 0.5 e 1, i termini per i quali l'indice assume valori più alti sono quelli con un contributo informativo maggiore per la descrizione del documento. Per evitare l'effetto dei singoli termini che possono presentarsi in maniera molto frequente nel documento, il *term frequency* può essere contenuto utilizzando una funzione del tipo $f(\mathrm{TF})$, ad esempio $\sqrt{(\mathrm{TF})}$ o $1 + \log(\mathrm{TF})$, così da assegnare una importanza relativa ai termini più frequenti.
Al contempo per poter valutare il livello di discriminazione dei termini all'interno della raccolta analizzata è opportuno introdurre un altro indice, l'*inverse document frequency* [77]. Indicando con df_i il numero di documenti in cui si presenta l'*i-mo* termine, l'espressione più diffusamente utilizzata per il calcolo dell'IDF è:

$$\mathrm{idf}_i = \log\left(\frac{n}{\mathrm{df}_i}\right) \tag{1.4}$$

dove n è il numero totale dei documenti nella collezione; l'uso del logaritmico è giustificato dalla necessità di compensare l'effetto del TF.
Il TF e l'IDF possono essere combinati in diversi modi per ottenere l'indice completo. Non esiste infatti una formulazione univoca e molto è lasciato all'esperienza del ricercatore e alla validità degli esperimenti empirici condotti negli anni [63]. Una delle possibili formulazioni è il *best fully weighted system*:

$$\frac{f_{ij} \cdot \log(n/\mathrm{df}_i)}{\sqrt{\sum_i \left[f_{ij} \cdot \log(n/\mathrm{df}_i)\right]^2}} \tag{1.5}$$

dove l'espressione al denominatore normalizza l'indice considerando la lunghezza del documento considerato.
Il principale vantaggio dello schema di ponderazione così costruito è la possibilità di confrontare i risultati ottenuti da *corpora* differenti. La 1.5 è senz'altro la formulazione più indicata quando è necessario adottare due sistemi di pesi differenti, come accade nei processi di *Information Retrieval*, dove è necessario distinguere il sistema di pesi utilizzato per i documenti dal sistema di pesi utilizzato per le *query*.
Scegliendo opportunamente un valore soglia, il TF/IDF può essere utilizzato per la selezione dei termini rilevanti ai fini dell'analisi successiva, rientrando così anche nell'ambito delle tecniche di *feature selection*.

1.1.4 Feature Selection

Quando si analizzano grandi raccolte di documenti con tecniche statistiche, il problema centrale è di certo l'elevata dimensionalità del vocabolario. Ogni singolo termine ne rappresenta una dimensione, quindi il vocabolario raggiunge facilmente decine o anche centinaia di migliaia di dimensioni. Si pensi ad esempio a quante parole sono utilizzate per scrivere una singola frase. Infatti, la matrice lessicale ($documenti \times termini$), costruita a seguito delle procedure di pulizia del testo, è una matrice sparsa e di dimensioni molto elevate. In realtà, come detto, solo una parte dei termini è effettivamente rilevante per determinare il contenuto dei documenti nella raccolta. Risulta, quindi, necessario eliminare il rumore presente nei dati, che rende il risultato di un'analisi inaffidabile e aumenta in modo significativo la complessità computazionale. Analogamente a quanto accade per l'analisi di grandi basi di dati numerici, l'approccio più comune per affrontare questo problema è procedere ad una selezione delle variabili, facendo riferimento a

tecniche di *feature selection*.
Nell'ambito dell'analisi dei dati testuali l'approccio più comune è l'utilizzo di una lista di *stop-word*. In generale questo elenco, detto *stoplist*, è una lista di parole che si presentano in una raccolta di documenti in maniera molto frequente e che non sono discriminanti del contenuto dei documenti stessi. Si tratta di parole che possiamo definire parti del discorso (*Part of speech* - POS) *funzionali*, poiché sono caratterizzate dal fatto di avere, all'interno di una grammatica un ruolo e un utilizzo definito, come ad esempio preposizioni, pronomi e congiunzioni, avverbi. Sono quindi strumentali alla costruzione di un testo e rappresentano l'elemento di congiunzione tra i termini che definiamo parti del discorso *lessicali*, quali sostantivi, aggettivi e verbi. Una *stop-list* può comprendere, oltre alle POS funzionali, termini molto specifici dell'ambito di indagine, perchè strettamente legati all'obiettivo dell'analisi. Ad esempio la parola "computer" può essere un termine discriminante in una raccolta di documenti non specifici, diventa invece una *stop-word* in un *corpus* dove tutti i documenti trattano di informatica.
In letteratura sono disponibili diverse proposte per la rimozione delle *stop-word* [52], poiché, soprattutto nell'ambito di applicazioni non supervisionate, l'utilizzo di una *stop-list* standard fornisce comunque un insieme valido di termini da eliminare.
Ciò nonostante, per quantificare in maniera più stringente l'importanza dei diversi termini nel *corpus* sono stati proposti una serie di metodi, alcuni dei quali specifici per tecniche supervisionate e riadattati al caso non supervisionato. Ne sono esempi misure quali il Term Strength [59], il Term Contribution [45] e l'Entropy-based Ranking [14]. Queste misure si basano sostanzialmente su funzioni di similarità (si veda a tal proposito il paragrafo 2.3), rendendo possibile la selezione dei termini rilevanti in maniera auto-

matica, senza avere necessariamente a disposizione informazioni *a priori*. É chiaro che la scelta della funzione di similarità comporta la possibilità di ottenere differenti risultati per la selezione dei termini rilevanti. Inoltre a causa dell'elevata dimensionalità del vocabolario di partenza, questi metodi sono caratterizzati da un'alta complessità computazionale, che in alcuni casi può essere superata facendo riferimento a tecniche di campionamento [14].

1.2 Riduzione della dimensionalità e visualizzazione dei dati testuali

Le tecniche di *feature selection* consentono, a partire dalla tabella lessicale T, di ridurre la dimensionalità del vocabolario. Tale riduzione, come detto, avviene attraverso la selezione di un sottoinsieme di termini del vocabolario, considerati rilevanti ai fini dell'analisi.
L'alternativa a questo approccio è ridurre l'elevata dimensionalità dello spazio definito dai termini del vocabolario attraverso l'impiego di metodi di *feature trasformation*. L'idea di fondo di questo tipo di metodi è operare una riduzione dimensionale definendo nuove caratteristiche che siano una rappresentazione funzionale delle caratteristiche dell'insieme dei dati d'origine.
I metodi di riduzione dimensionale, che hanno trovato largo impiego nell'ambito dell'analisi dei dati testuali, sono i metodi fattoriali [54], che, attraverso una riduzione del numero di variabili del fenomeno, producono delle nuove variabili sintetiche in grado di ricostruire i principali assi semantici che caratterizzano la variabilità dei contenuti del *corpus*.
Con il ricorso a tecniche di tipo fattoriale i documenti vengono rappresentati in un spazio di dimensioni ridotte, dove le nuove dimensioni sono sostan-

zialmente una combinazione lineare dei termini nel dataset di partenza.
Metodi come il *Latent Semantic Indexing* (LSI) [18] e l'*Analisi delle Corrispondenze Lessicali* (ACL), pur essendo stati sviluppati in contesti e con obiettivi diversi, si basano su questo comune principio.
Il LSI nasce nell'ambito degli strumenti di *Information Retrieval*, allo scopo di individuare ed indicizzare i documenti rilevanti a fronte di una richiesta dell'utente (query). Il principale vantaggio di questo metodo consiste nel superare i limiti dell'effettiva presenza, nei documenti, di tutte le parole chiave utilizzate nella ricerca (query). Partendo quindi da una rappresentazione vettoriale dei documenti è possibile determinare uno spazio di concetti "artificiali" analizzando se e quanto di frequente determinati termini vengono usati insieme, facendo emergere le relazioni semantiche "latenti" tra termini e documenti.
L'ACL è una delle tecniche di Analisi Multimensionale dei Dati, il cui obiettivo è, a partire dalla tabella lessicale T, descrivere da un punto di vista geometrico e algebrico delle relazioni tra i termini, tra i documenti e, indirettamente, tra termini e documenti. L'analisi è svolta calcolando una serie di fattori a partire dalle variabili originarie, ognuno dei quali rappresenta una dimensione latente del tipo di associazione presente nei dati. La successiva rappresentazione in forma grafica consente una interpretazione semplice della struttura, evidenziando gli aspetti non rilevabili dalla lettura diretta dei dati.
Il problema della ricerca di un sottospazio di riferimento per l'individuazione di una struttura di associazione tra le variabili osservate è, quindi, di importanza fondamentale tanto per le tecniche di Analisi Multidimensionale dei dati, quanto nelle strategie di *Information Retrieval*.
Questo si risolve ricorrendo alla *decomposizione in valori singolari* (DVS)

[24], attraverso la quale è possibile decomporre la matrice dei dati originaria e ricostruirla come matrice di rango ridotto.
Data una matrice rettangolare A (n, p) con $n > p$ di rango p, si ha che:

$$\begin{aligned} \mathrm{A}&=\mathrm{U}\Lambda V^T \\ \mathrm{U}^T U &= V^T V = I \end{aligned} \tag{1.6}$$

dove Λ (p, p) è una matrice diagonale di numeri positivi λ_α (con $\alpha = 1, 2 \ldots p$) in ordine decrescente detti valori singolari; $U(n, p)$ e $V(p, p)$ sono le matrici dei vettori singolari di sinistra e destra.
In Balbi e Misuraca [2] si sottolinea come una rilettura della decomposizione in termini di DVS *generalizzata* (DVSG) [32] consenta di ricondurre il problema della scelta del sistema di ponderazione per termini e documenti ad un problema di definizione di metriche Euclidee ponderate in spazi multdimensionali.
Nella DVSG l'individuazione di un sottospazio che meglio approssimi la struttura dei dati è espressa in maniera equivalente da:

$$\begin{aligned} \Omega^{\frac{1}{2}} A \Phi^{\frac{1}{2}} &= U\Lambda V^T \\ \mathrm{U}^T U = V^T V &= I \end{aligned} \iff \begin{aligned} \mathrm{A}&=\mathrm{U}\Lambda V^T \\ \mathrm{U}^T \Omega U &= V^T \Phi V = I \end{aligned} \tag{1.7}$$

dove Ω (n, n) e Φ (p, p) sono due matrici simmetriche definite positive che rappresentano i sistemi di pesi e le metriche Euclidee ponderate scelte nei due spazi di rappresentazione degli elementi posti sulle righe e sulle colonne della matrice A.
Metodi di riduzione dimensionale come il LSI e l'ACL si differenziano, quindi, rispetto al sistema di pesi e alle metriche in cui sono proiettati i termini e i documenti. Concordemente con gli obiettivi dell'analisi si prediligerà una tecnica piuttosto che un'altra. Se infatti lo scopo dell'analisi è quello

di ricercare una conoscenza specifica, allora il LSI consente di enfatizzare i termini più presenti e i documenti più lunghi. Se lo scopo dell'analisi è quello di una conoscenza più ampia in un'ottica esplorativa allora metodi di Analisi Multidimensionale dei Dati, come l'ACL forniscono dei risultati più interessanti.

Capitolo 2

Il Clustering di documenti in linguaggio naturale

L'affermazione riportata da Tyron e Bailey nel loro lavoro del 1970 [82]:

"Understanding our world requires conceptualizing the similarities and differences between the entities that compose it"

esprime perfettamente la naturale tendenza umana a classificare un insieme di oggetti in gruppi omogenei e quindi ad identificarli con specifiche categorie. La classificazione agevola l'interpretazione della realtà fenomenica, rappresentando un momento essenziale del procedimento scientifico, e per tale ragione accomuna i più svariati campi di ricerca: le Scienze Sociali come la Psicologia, la Statistica, la Biologia, l'Informatica, e così via.
Lo sviluppo di tecniche rivolte alla selezione e al raggruppamento di oggetti omogenei in un dato insieme è stato, quindi, una vera e propria sfida interdisciplinare, e ha dato luogo ad una letteratura molto vasta e ancor oggi in

crescita, ma troppo spesso poco integrata. A tal proposito basti pensare, ad esempio, alla differente terminologia che si incontra nel passare da una disciplina ad un'altra, cosa che rende spesso difficile lo scambio tra ricercatori che lavorano prettamente in ambiti diversi: in Statistica ci si riferisce a tale famiglia di metodi come *Cluster Analysis* o Analisi dei gruppi, in Biologia si usa spesso il termine *Numerical Taxonomy*, nelle Scienze Sociali quello di *Tipologia*, nell'ambito del Data Mining ci si riferisce alla *Cluster Analysis* con il termine *Apprendimento non supervisionato*.

Le ragioni di questo interesse sono sostanzialmente due: la polivalenza di questo strumento in diversi contesti applicativi, la necessità di far fronte all'ingente mole di dati che ha accompagnato lo sviluppo tecnologico. Se da un lato sono state proposte diverse metodologie di *Clustering* per trattare dati dalle caratteristiche più disparate (misure, attributi, testi, immagini), con il fine ultimo di individuare lo schema di raggruppamento che ne rifletta sostanzialmente la naturale struttura sottostante, dall'altro nel tempo si è posta maggiore attenzione agli aspetti algoritmici e computazionale di queste tecniche. Pertanto, da ciascuno dei metodi di *Clustering*, che si distinguono rispetto al criterio (da ottimizzare) in base al quale sono individuati i raggruppamenti, sono derivati una serie di algoritmi, realizzati per rendere il processo di classificazione più efficiente in termini di calcolo e per essere in grado di trattare dati altamente dimensionali.

K. Pearson fu il primo che, sul finire del secolo XIX, affrontò il problema della classificazione da un punto di vista statistico, anche se il termine *Cluster Analysis* fu utilizzato per la prima volta da Robert Tryon [81] solo nel 1939, presentando la propria teoria e inquadrandola come una variante dell'analisi fattoriale. A partire dagli anni sessanta la *Cluster Analysis* riceve un forte impulso grazie sia al lavoro di due biologi, Sokal e Sneath [76], sia al

lavoro di Ward [85], che nel 1963 elabora una propria tecnica di *Clustering* a partire da un problema di classificazione di posizioni occupazionali. Nel 1967 Mac Queen [58] propone il metodo di *Clustering* non gerarchico che ad oggi risulta ancora il più ampiamente utilizzato: il *K-means*. Il successo di questo metodo è dovuto alla sua semplicità e ai suoi vantaggi computazionali, che lo rendono in grado di trattare grandi moli di dati [44].
Da allora ai giorni nostri, i contributi scientifici in questo settore di ricerca si sono moltiplicati. Limitandoci unicamente alla terminologia utilizzata in Statistica ed effettuando una ricerca nella letteratura accademica su Google Scholar, gli articoli che contengono il termine *Cluster Analysis* al 2015 sono circa 900.000.

2.1 Formalizzazione del problema

Il problema di base è pressoché identico in tutti i diversi contesti di utilizzo e può essere formulato come segue: partendo da un collettivo multidimensionale si assegnano le unità (individui, oggetti, eventi, testi, immagini, ecc.) a categorie non definite *a priori*, formando dei gruppi (i *cluster*), tali che le unità che vi appartengono siano tra loro simili, rispetto all'insieme di caratteristiche prese in considerazione e secondo uno specifico criterio.
Consideriamo un insieme di n oggetti diversi:

$$O = (o_1, \ldots, o_n)$$

dove ciascun oggetto o_i (per $1 \leq i \leq n$) è definito da p valori cui possiamo riferirci come caratteristiche o condizioni. Da ciò deriva che ogni i-esimo oggetto può essere considerato come un elemento appartenente ad uno spazio

p-dimensionale. Definita con $C_k = (c_1, \ldots, c_k)$ una partizione di O, cioè un insieme di parti di O tali che $\bigcup_{i=1}^{k} c_i = O$ e $c_i \cap c_j = \emptyset$ (per $1 \leq i \neq j \leq k$), possiamo indicare ciascun c_i come un *cluster* di oggetti e quindi C_k come una possibile soluzione di Clustering.
Lo scopo della *Cluster Analysis* è quello di determinare una partizione C_k dell'insieme O in base ad una misura di similarità o distanza M, definita sugli elementi dell'insieme medesimo. In particolare vale la regola generale per cui si vuole che gli oggetti appartenenti ad uno stesso cluster abbiano tra loro la massima similarità, mentre gli oggetti di cluster diversi siano quanto più dissimili.

2.2 Le fasi di una *Cluster analysis*

Un processo di *Clustering* si articola in alcune fasi fondamentali e richiede una serie di scelte, tra le quali quella di uno specifico algoritmo di raggruppamento. É necessario considerare compiutamente diversi aspetti [30]:

- la scelta delle caratteristiche in base alle quali raggruppare gli oggetti;
- la scelta di una adeguata misura della dis/somiglianza esistente fra gli oggetti;
- la scelta dell'algoritmo di raggruppamento;
- la validazione dei risultati ottenuti.

Tali scelte possono condizionare notevolmente la soluzione di *Clustering* ottenuta, che potrebbe essere di difficile interpretazione o ancor peggio potrebbe non riflettere la naturale struttura dei dati.

Nei seguenti paragrafi saranno illustrate le alternative, proposte in letteratura, sulla base delle quali scegliere una strategia che sia coerente con l'obiettivo dell'analisi.

2.2.1 Scelta delle caratteristiche

La quantità e la varietà di dati ad oggi disponibile è aumentata esponenzialmente nel tempo. Tuttavia, qualunque ne sia il volume e la natura, per poter essere analizzati con tecniche statistiche è necessario organizzarli in una matrice di dati, dove ciascuna riga rappresenta un'unità, descritta da una serie di attributi o valori che la caratterizzano, manifestazioni di determinate variabili. Con riferimento al contesto applicativo, il termine unità potrà essere sostituito da entità, istanze, record, esempi, punti, oggetti, transazioni, *feature-vectors*, e così via. Allo stesso modo le variabili potranno essere chiamate condizioni, proprietà, dimensioni, *feature*.
Formalmente, la classica rappresentazione di una matrice di dati X unità $\times$ variabili $(n \times p)$ è la seguente:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

dove x_{ij} rappresenta l'attributo/valore della j-esima caratteristica osservato sulla i-esima unità. A partire da una struttura dati di questo tipo, il primo passo di una *Cluster Analysis* è scegliere se si intende raggruppare le righe o le colonne della matrice dei dati, e rispetto a quali caratteristiche ricercare il raggruppamento. In letteratura [] sono stati comunque proposti metodi per operare contemporaneamente su righe e colonne, se ciò dovesse

essere interessante rispetto al particolare problema (si parla generalmente di *co-clustering*).
Se si sceglie di raggruppare le unità l'obiettivo è individuare nei dati *pattern* distintivi che possano identificare dei gruppi con caratteristiche distintive. La *Cluster Analysis* potrebbe per contro essere vista come uno *step* di un processo di analisi più articolato, con la funzione di operare una riduzione dimensionale dei dati agendo sulle colonne della matrice.
La scelta delle variabili, rispetto alle quali raggruppare le unità, è naturalmente legata agli obiettivi dell'analisi. Dalla matrice di partenza devono infatti essere selezionate quelle variabili che si ritengono significative per l'identificazione dei *cluster*. Inoltre, ad oggi, fenomeni tipici di settori quali la medicina, la biologia, la genetica, la finanza, l'astronomia, l'informatica, sono caratterizzati da un numero estremamente elevato di variabili, quindi la loro rappresentazione raggiunge talvolta dimensioni considerevoli. Per poter affrontare allora il problema diventa cruciale la selezione di un numero più contenuto ma comunque sufficiente di variabili. La riduzione della dimensionalità può essere ottenuta sia attraverso tecniche trasformazione delle variabili (*feature trasformation*) [50], sia attraverso metodi di selezione delle variabili (*feature selection*) [10].
Le tecniche di *feature trasformation* si basano sulla rappresentazione dei dati in uno spazio ridotto, preservando generalmente la distanza relativa originaria tra le unità. La sintesi dei dati è ottenuta creando combinazioni lineari delle variabili, attraverso le quali è possibile evidenziare strutture latenti nei dati. Tuttavia, tali tecniche non rimuovono dall'analisi nessuno degli attributi originari, e inoltre gli attributi trasformati sono spesso difficili da interpretare e ciò può rendere i risultati del *Clustering* meno utili. Pertanto la trasformazione degli attributi è adottata soltanto per insiemi di

dati in cui gran parte delle dimensioni sono rilevanti per l'attività di *Clustering*.
Nella pratica accade spesso che alcuni degli attributi disponibili siano rilevanti solo rispetto a certi insiemi di unità. Ciò vuol dire che i gruppi possono esistere in sottospazi differenti dello stesso insieme dei dati. Le tecniche di *feature selection* comportano la ricerca di appropriati sottoinsiemi di variabili per descrivere la somiglianza delle unità che appartengono allo stesso gruppo, e la valutazione degli stessi sottoinsiemi utilizzando opportuni criteri.
La scelta tra questi due approcci per la selezione delle variabili è determinante per l'efficacia di un'applicazione di *Clustering*. Una "elegante" selezione delle variabili può ridurre notevolmente il carico di lavoro e semplificare le successive fasi dell'analisi. In linea generale, una variabile "ideale" dovrebbe essere in grado di distinguere l'appartenenza delle unità a gruppi distinti, essere cioè immune al rumore, semplice da estrarre e soprattutto da interpretare.

2.2.2 Scelta di un criterio di valutazione della somiglianza

L'obiettivo della *Cluster Analysis* come detto è quello di suddividere un collettivo, eterogeneo al suo interno, in un certo numero di gruppi secondo il livello di dis/somiglianza tra le unità che lo compongono.
É naturale chiedersi che tipo di misure possono essere utilizzate per determinare la somiglianza (o dissomiglianza), o in altri termini, come misurare la distanza tra coppie di unità, un'unità e un *cluster*, o coppie di *cluster*. Ovviamente, la scelta di una misura di dis/somiglianza influisce direttamente sulla formazione dei *cluster*, in quanto le diverse misure ne sottendono specifiche definizioni in termini di forma, dimensione e densità.

Per poter applicare molti degli algoritmi di *Clustering*, è necessario trasformare la matrice dei dati originaria in una matrice del tipo seguente:

$$P = \begin{bmatrix} p_{11} & \dots & p_{1n} \\ p_{21} & \dots & p_{2n} \\ \dots & \dots & \dots \\ p_{n1} & \dots & p_{nn} \end{bmatrix}$$

Si tratta di una matrice simmetrica $n \times n$, dove il generico elemento p_{ij} esprime la prossimità tra l'unità i e l'unità j, con $1 \leq i \neq j \leq n$.
Tale rappresentazione dei dati può risultare da considerazioni soggettive sulle differenze tra le unità, così come da calcoli effettuati sulla matrice stessa. In tal caso il criterio da adottare è strettamente legato tanto al tipo di variabili coinvolte, quanto al contesto di riferimento. Le variabili possono essere distinte in variabili qualitative, quantitative o miste.
Nel caso di variabili qualitative si parla generalmente di similarità, poiché la valutazione sulle caratteristiche delle unità può essere fatta solamente in termini di uguaglianza/disuguaglianza. Se consideriamo ad esempio due individui potremo dire che sono simili o dissimili a seconda che condividano o meno determinate caratteristiche socio-demografiche quali titolo di studio, professione, e così via. In tali casi quindi la somiglianza tra due unità i e j è valutata attraverso funzioni (misure) di similarità s_{ij}, elemento generico di una matrice S $(n \times n)$, che confrontano i vettori p-dimensionali x_i e x_j.
Una funzione di similarità in generale deve soddisfare le seguenti proprietà:

1. $0 \leq s_{ij} \leq 1 \quad \forall\, 1 \leq i \neq j \leq n$
2. $s_{ii} = 1$ (una unità è sempre massimamente simile a sé stessa)
3. $s_{ij} = s_{ji}$ (la similarità tra due unità è biunivoca)

Se inoltre soddisfa la seguente condizione parleremo di metrica di similarità:

$$s_{ij}s_{jk} \leq [s_{ij} + s_{jk}]s_{ik} \quad \forall \ 1 \leq i \neq j \neq k \leq n$$

Per variabili quantitative la somiglianza tra due unità può essere espressa non solo in termini di uguaglianza/disuguaglianza, ma è possibile anche calcolare lo scarto tra i valori osservati per le diverse unità e quindi rappresentare l'eventuale diversità in termini di distanza. In tal caso la matrice ricavata a partire dai dati originari può essere indicata con D, il cui generico elemento d_{ij} rappresenta la distanza tra due vettori p-dimensionali x_i e x_j. La scelta tra le diverse formulazioni della funzione di distanza è dipendente dal problema in esame. Una misura di distanza deve soddisfare comunque sempre le seguenti proprietà:

1. $d_{ij} \geq 0$ (la distanza è sempre non negativa)
2. $d_{ii} = 0$ (la distanza di una unità da se stessa è nulla)
3. $d_{ij} = d_{ji}$ (la distanza tra due unità è simmetricamente biunivoca)

Una misura di distanza è chiamata poi metrica se soddisfa oltre alle condizioni su elencate anche la cosiddetta disuguaglianza triangolare:

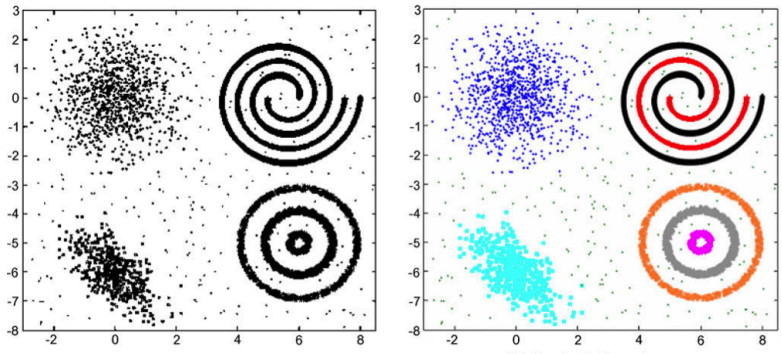
$$d_{ij} \leq d_{ik} + d_{kj} \quad \forall \ 1 \leq i \neq j \neq k \leq n$$

Una rassegna esaustiva delle diverse misure di distanza e di similarità comunemente utilizzate è disponibile ad esempio in [30]. Per poter trattare congiuntamente variabili di natura diversa, qualitative e quantitative, si devono ridurre alla scala di precisione inferiore le seconde. Per evitare tale perdita di informazione si può utilizzare ad esempio l'indice di Gower [31].

2.2.3 Scelta di un algoritmo di raggruppamento

In generale, un algoritmo di *Clustering* nasce dalla combinazione della scelta di una misura di prossimità/distanza tra le unità da raggruppare e di della scelta di una funzione criterio in base al quale effettuare il raggruppamento. Definendo in che modo misurare il livello di somiglianza tra le unità si decide implicitamente il "tipo di gruppi" da cercare nei dati. A partire poi da una definizione di cosa costituisce un gruppo (*cluster*), un algoritmo di *Clustering* necessita della definizione della funzione criterio, allo scopo di individuare il "miglior" raggruppamento dei dati che rifletta tale struttura. Il problema del *Clustering* si traduce, quindi, in un problema di ottimizzazione, matematicamente ben definito e che trova diverse soluzioni in letteratura [21].
Dalle diverse combinazioni di questi due aspetti sono stati sviluppati un gran numero di algoritmi, rivolti alla soluzione di problemi diversi in specifici campi di ricerca. Tuttavia, a fronte di tutte queste possibili soluzioni che perseguono un medesimo obiettivo, non esiste un algoritmo che sia in grado di risolvere universalmente tutti i problemi, come dimostrato nel *teorema dell'impossibilità* di Kleinberg [48].
Si osservi il dataset in Fig. 2.1(a), e si supponga di voler individuare in maniera automatica il raggruppamento naturale dei dati mostrato nella Fig. 2.1(b). Nell'esempio i *cluster* si differenziano per forma, dimensione, e densità, e sebbene possano essere evidenti per un operatore umano, nessuno tra gli algoritmi disponibili è in grado di individuarli tutti. Inoltre all'aumentare del numero di dimensioni (più di tre dimensioni), nemmeno un essere umano è in grado di individuare i possibili raggruppamenti dei dati. Diventa quindi impellente la necessità di utilizzare algoritmi automatici che siano

in grado di trattare dati, caratterizzati da un elevato numero di descrittori.



(a) Dataset di partenza (b) Raggruppamento desiderato

Figura 2.1: Esempio di *Clustering*

Pertanto, per selezionare o progettare una strategia di *Clustering* appropriata è importante esaminare attentamente le caratteristiche del problema, tenendo conto delle ipotesi, implicite nel metodo, sulla forma dei *cluster* o sulla struttura di partizionamento, basate sulle misure di prossimità/distanza e sulla funzione criterio. Nel paragrafo 2.4 è presentata una panoramica dei principali metodi di *Clustering*, utilizzati specificatamente per il trattamento di *corpora* testuali.

2.2.4 Validazione e interpretazione dei risultati

La validazione della struttura di partizionamento e l'effettiva interpretabilità dei *cluster* risultanti rappresentano un aspetto fondamentale della *Cluster Analysis*, di cui spesso non si tiene conto. L'insieme di queste tecniche è infatti connotato come parte del più ampio contenitore dei metodi di tipo

esplorativo, in quanto possono mettere in evidenza associazioni e strutture nei dati non altrimenti rilevabili. In tal senso la validazione dei risultati può rappresentare un'aggiunta onerosa a ciò che può essere considerato un processo informale. Tuttavia, dato un insieme di dati, ogni algoritmo di *Clustering* ne genera una suddivisione anche quando i dati non presentano alcun raggruppamento naturale. Ne consegue che il successo delle applicazioni di *Cluster Analysis* dipende completamente dal fatto di sapere se il modello di raggruppamento imposto corrisponde ad una struttura reale o meno. Inoltre, differenti approcci conducono spesso a soluzioni diverse; e anche per lo stesso algoritmo, la scelta dei parametri di *input* o addirittura l'ordine delle unità nel dataset possono influenzare i risultati finali.
L'uso di criteri di valutazione efficaci è importante per fornire all'utente risultati con un certo grado di affidabilità. Tali valutazioni dovrebbero essere obiettive e indipendenti dall'algoritmo scelto per il raggruppamento. Inoltre, devono essere utili per individuare il numero di *cluster* presenti nei dati, per valutare se i *cluster* ottenuti sono significativi o sono solo un artefatto degli algoritmi, o ancora per decidere quale tra i diversi algoritmi scegliere. I diversi aspetti della validazione saranno discussi in maniera dettagliata nel Capitolo 3.

2.3 Misure di similarità e distanza tra i Documenti

In molti dei task tipici del *Text Mining*, e ancor più nel *Clustering*, risulta di fondamentale importanza stabilire dei criteri per poter valutare il diverso livello di somiglianza tra i documenti. Come detto in precedenza gli algoritmi di *Clustering*, utilizzati in molti ambiti diversi così come in generale in un quadro di analisi statistica, sono basati sul concetto di vicinanza o di

separazione tra le unità osservate. Se tali concetti sono di semplice e spesso intuitiva formulazione per i tradizionali dati quantitativi e qualitativi, diventa molto più complesso riportare la loro definizione al trattamento dei dati testuali, e quindi dei documenti.
In un approccio di tipo qualitativo è possibile leggere documenti differenti per natura e contenuto, individuare i temi principali e i concetti chiave, e quindi formulare un giudizio sul grado di somiglianza. La comprensione della somiglianza, espressa in termini di significato e informazione, diventa però di difficile portata in tutte quelle strategie quantitative (automatiche) che devono necessariamente prevedere delle semplificazioni per poter essere implementate.
L'unica possibile soluzione è quella di "forzare" misure di similarità e di distanza proposte in contesti differenti per poter valutare la somiglianza di documenti in termini di contenuto informativo.

2.3.1 La similarità

La codifica dei documenti attraverso lo schema introdotto nel paragrafo 1.1.2 riconduce il problema del loro confronto alle misure di similarità tra vettori. I documenti, come visto, possono essere rappresentati ricorrendo ad uno schema di ponderazione *booleano*, nel quale si assegna 0 ai termini del vocabolario non presenti in uno specifico documento della collezione e 1 ai termini presenti (quale che sia la loro occorrenza). Dati due documenti D_1 e D_2 la più semplice misura di similarità da utilizzare, in termini computazionali, è il cosiddetto *coefficiente di matching*:

$$s_M(D_1, D_2) = |D_1 \cap D_2| \equiv \sum_{i=1}^{p} w_{i1} \cdot w_{i2} \tag{2.1}$$

Tale misura tiene conto di quanti valori non nulli sono presenti in entrambi i vettori, ma ha il limite di non considerarne l'eventuale diversa dimensione. Se si considera il *matching* di due documenti ciò si traduce rispettivamente nel valutare quanti termini sono presenti in entrambi senza però tenere conto della lunghezza dei documenti stessi. Per ovviare a tale problema si può ricorrere in alternativa al *coefficiente di Dice*, che a differenza del matching è normalizzato per la dimensione:

$$s_D(D_1, D_2) = \frac{2\,|D_1 \cap D_2|}{|D_1| + |D_2|} \equiv \frac{2\sum_i w_{i1} \cdot w_{i2}}{\sum_i w_{i1} + \sum_i w_{i2}} \tag{2.2}$$

Essendo una misura normalizzata consente una lettura dei valori ottenuti in un range compreso tra 0 (max dissimilarità) e 1 (max similarità), confrontabili quindi se riferiti a documenti differenti. Il *coefficiente di Jaccard* è un'altra misura di similarità spesso utilizzata in questo come in molti altri domini differenti:

$$s_J(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} \equiv \frac{\sum_i w_{i1} \cdot w_{i2}}{\sum_i w_{i1} + \sum_i w_{i2} - \sum_i w_{i1} \cdot w_{i2}} \tag{2.3}$$

Formalmente vengono comparati il numero di valori non nulli contemporaneamente presenti in due vettori rispetto al numero di valori non nulli presenti singolarmente in ciascun vettore. Ciò implica che si tiene conto di quanti termini sono condivisi da due documenti rispetto a quelli presenti nell'uno o nell'altro. Tale misura "penalizza" ovviamente i documenti con un numero di termini comuni relativamente piccolo in proporzione a tutti i termini presenti, ma ha al contempo il vantaggio di fornire valori più bassi rispetto alle altre misure proposte per i casi di bassa sovrapposizione (*low-overlap*) tra documenti. Anche in questo caso i valori sono compresi tra 0

e 1.
Se si ritiene più utile una rappresentazione dei termini presenti nei documenti in base al numero di occorrenze (quindi uno schema di ponderazione *frequentista*), allora il *coseno* rappresenta indubbiamente una misura di similarità di estremo interesse, soprattutto per il particolare significato che esso riveste all'interno di uno spazio vettoriale, come quello p-dimensionale nel quale i documenti sono rappresentati:

$$s_C(D_1, D_2) = \frac{|D_1 \cap D_2|}{\sqrt{|D_1| \times |D_2|}} \equiv \frac{\sum_i w_{i1} \cdot w_{i2}}{\sqrt{\sum_i w_{i1}^2 \cdot \sum_i w_{i2}^2}} \tag{2.4}$$

Da un punto di vista statistico il coseno può anche essere interpretato in termini di correlazione lineare. Il valore varia tra 0 se i due vettori sono ortogonali (massima dissimilarità tra i documenti) e 1 se i vettori hanno la stessa direzione (massima similarità tra i documenti). È interessante notare come tale misura sia comunque indipendente dalla dimensione del vettore; riportando ancora una volta il problema ai documenti, ciò implica come ad esempio risultino simili documenti con gli stessi termini ma con lunghezza diversa.

2.3.2 Dalla similarità alla distanza

In generale se le variabili considerate sono quantitative la somiglianza tra due unità può essere espressa non solo in termini di uguaglianza/disuguaglianza, ma è possibile anche calcolare lo scarto tra i valori osservati per le diverse unità e quindi rappresentare l'eventuale diversità in termini di distanza. Se i documenti sono rappresentati seguendo uno schema di ponderazione più complesso come il *Term frequency/Inverse document frequency*,

dove il peso attribuito ad ogni termine nei diversi documenti è assimilabile ad un'intensità, è quindi possibile fare riferimento a misure di distanza.
Una espressione generale della distanza tra due vettori documento D_1 e D_2, in forma matriciale, è data da:

$$d(D_1, D_2) = \sqrt{(D_1 - D_2)' M (D_1 - D_2)} \tag{2.5}$$

dove la metrica M è una matrice simmetrica definita positiva.
A seconda del tipo di metrica utilizzata nella 2.5 vengono definite differenti misure di distanza. Se si considera $M \equiv I$ con I matrice unitaria, si ottiene allora la *distanza Euclidea*:

$$d_E(D_1, D_2) = \sqrt{(D_1 - D_2)'(D_1 - D_2)} \tag{2.6}$$

La *distanza Euclidea* è stata particolarmente usata nell'ambito del *Clustering* soprattutto per la sua semplice interpretazione geometrica, è infatti alla base di molti algoritmi, tra i quali il Kmeans. Tuttavia il maggior limite della *distanza Euclidea* in un contesto di *Text Mining* è che talvolta due documenti D_1 e D_2 possono risultare molto simili, pur non avendo termini in comune.
Se si utilizza come metrica $M \equiv \Sigma$, con Σ matrice di varianza-covarianza, si ha la cosiddetta *distanza di Mahalanobis*:

$$d_M(D_1, D_2) = \sqrt{(D_1 - D_2)' \Sigma^{-1} (D_1 - D_2)} \tag{2.7}$$

La *distanza di Mahalanobis* consente di "eliminare" la correlazione tra le variabili ed equivale di conseguenza ad una *distanza Euclidea* calcolata su variabili standardizzate.
Nel caso in cui si voglia attribuire diversa importanza alle variabili considerate è possibile utilizzare delle distanze ponderate, anche se la determinazione dei pesi lascia ampi margini alla soggettività. Un particolare

tipo di distanza, di fatto una metrica euclidea *ponderata*, è la distanza del *Chi-quadro*, basata sulla statistica χ^2. Nelle tabelle di contingenza, la generica cella in corrispondenza della i-ma modalità del carattere I e della j-ma modalità del carattere J contiene la frequenza f_{ij}, cui corrispondono le frequenze marginali $f_{i.}$ e $f_{.j}$, rispettivamente di riga e di colonna. La metrica del Chi-quadro definisce la distanza tra due righe o tra due colonne come:

$$d^2(i, i') = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \tag{2.8}$$

$$d^2(j, j') = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{2.9}$$

Rispetto ad altre metriche, la distanza del Chi-quadro, che è alla base dell'Analisi delle Corrispondenze, gode dell'importante proprietà dell'*equivalenza distributiva*, risultando invariante rispetto ai criteri di codifica o al modo di aggregare le entità in gruppi, a condizione che le unità aggregate siano omogenee. É possibile considerare due punti molto prossimi, e quindi con un profilo simile, come un unico punto che abbia per massa la somma delle frequenze dei punti originari.
Il vantaggio/svantaggio dell'utilizzo di tale distanza è strettamente connesso al fatto che le modalità meno frequenti e quelle più frequenti, ponderate per il reciproco delle loro frequenze marginali, sono ugualmente ben rappresentate. Nel trattamento statistico delle basi documentali il dataset iniziale è filtrato a monte, perchè l'attenzione del ricercatore è rivolta principalmente all'analisi e alla successiva rappresentazione dei termini a maggior contenuto informativo. Pertanto è necessario considerare di volta in volta, a seconda dell'obiettivo dell'analisi, la convenienza dell'utilizzo di metodi basati sul

Chi-quadro.
É bene precisare che da un indice di similarità è possibile ricavare il corrispondente indice di *dissimilarità* come il complemento a 1; se viene soddisfatta anche la cosiddetta disuguaglianza triangolare (si veda paragrafo 2.2.2), allora l'indice di *dissimilarità* così individuato rappresenta una misura della distanza tra i documenti considerati.

2.4 I metodi di Clustering per i documenti

Sebbene la *Cluster Analysis* sia incentrata su un obiettivo intuitivamente convincente, questo risulta vagamente definito. I metodi sono accomunati solo rispetto a questo livello molto generale di descrizione, per poi differenziarsi rispetto al criterio specifico che ottimizzano e alla specifica definizione di *cluster* che implicitamente sottintendono. Una classificazione dei metodi di *Clustering* ampiamente condivisa in letteratura [43] [28] suggerisce di distinguere i metodi in due grandi famiglie, sulla base delle proprietà dei *cluster* generati:

- **metodi gerarchici** (*divisivi* o *agglomerativi*), in cui viene costruita una gerarchia di partizioni annidate caratterizzate da un numero (de)crescente di gruppi.

- **metodi partitivi**, in cui un insieme di unità viene suddiviso in un pre-specificato numero di gruppi, ottenendo così un'unica partizione dei dati.

Il continuo proliferare di contributi scientifici e la diversificazione degli studi in questo settore di ricerca ha comportato un aumento costante nel numero di algoritmi di *Clustering* e con esso il numero di famiglie. Oltre ai metodi

gerarchici e ai metodi partitivi sono solitamente considerate due ulteriori classi di metodi:

- **metodi basati sulla densità**, in cui i *cluster* sono identificati come regioni dello spazio dense, separate da zone a più scarsa densità, che rappresentano quindi il rumore. Questi metodi si caratterizzano per la capacità intrinseca di individuare *cluster* di forma arbitraria e di filtrare il rumore presente nei dati identificando gli *outlier*. Da queste considerazioni è lecito attendersi una buona qualità del risultato del *Clustering*, tuttavia questi metodi producono a volte risultati poco significativi e difficilmente interpretabili. Dal punto di vista teorico sarebbe conveniente poter disporre *a priori* dei valori corretti per i parametri che definiscono la densità dei *cluster*, ed eventualmente differenziarli per i diversi gruppi, ma nella pratica non è ragionevole pensare di ottenere questo tipo di informazione. Ciò comporta dei limiti oggettivi che costituiscono un problema di ricerca ancora aperto.

- **metodi basati sulla suddivisione dello spazio**, in cui si utilizza un approccio fondamentalmente diverso dai precedenti, sostanzialmente si ragiona sullo spazio piuttosto che sui dati. Lo spazio viene quantizzato in un numero finito di celle sulle quali viene effettuato il processo di *Clustering*. Questo tipo di metodi, noti anche come metodi *Grid-based*, consentono una veloce computazione indipendente dal numero di unità da classificare ma unicamente dipendente dal numero di celle in cui lo spazio viene quantizzato. Le celle riassumono il contenuto delle unità in esse, ciò rende possibile mantenere l'intero dataset in memoria centrale; inoltre operando sulle singole celle questi algoritmi sono intrinsecamente parallelizzabili. Sono inoltre in grado di rico-

noscere *cluster* di forma arbitraria e i risultati sono modestamente sensibili rispetto alla scelta dei parametri di ingresso. Il limite di questi metodi risiede nella qualità dei cluster risultanti che dipende in modo significativo da quanto fine è la quantizzazione effettuata, con la quale è quindi necessario stabilire un *trade-off*.

Tra le varie classificazioni dei metodi di *Clustering* comunemente utilizzate, una ulteriore categorizzazione dipende dalla possibilità che un'unità possa essere o meno assegnata a più gruppi:

- **metodi esclusivi** in cui ogni unità può essere assegnata ad uno e ad un solo gruppo. I *cluster* risultanti, quindi, non possono avere elementi in comune. Questo approccio, detto anche *Hard* o *Crisp*, è alla base della maggior parte dei metodi di *Clustering*.
- **metodi non esclusivi**, in cui un'unità può essere assegnata a più gruppi con un diverso grado di appartenenza. Questi metodi, noti come metodi di *Soft* o *Fuzzy Clustering*, possono condurre a schemi di raggruppamento maggiormente compatibili con le situazioni reali, gestendo in maniera più efficace l'incertezza dei dati reali.

Lo sviluppo dei metodi di *Clustering* precede la sua applicabilità nell'ambito del *Text Mining*. Infatti molti dei metodi in Fig. 2.2, come ad esempio i metodi basati sulla densità o sulla suddivisione dello spazio, sono stati sviluppati in altri contesti e definiscono il concetto di gruppo in una maniera non idonea per il *Clustering* di documenti.
Questo perché i documenti, rispetto ai classici dati numerici, hanno un insieme di proprietà uniche di cui non si può non tener conto. La tabella lessicale attraverso la quale si rappresenta l'informazione testuale contenuta nei documenti raggiunge dimensioni molto elevate ed è molto sparsa, in

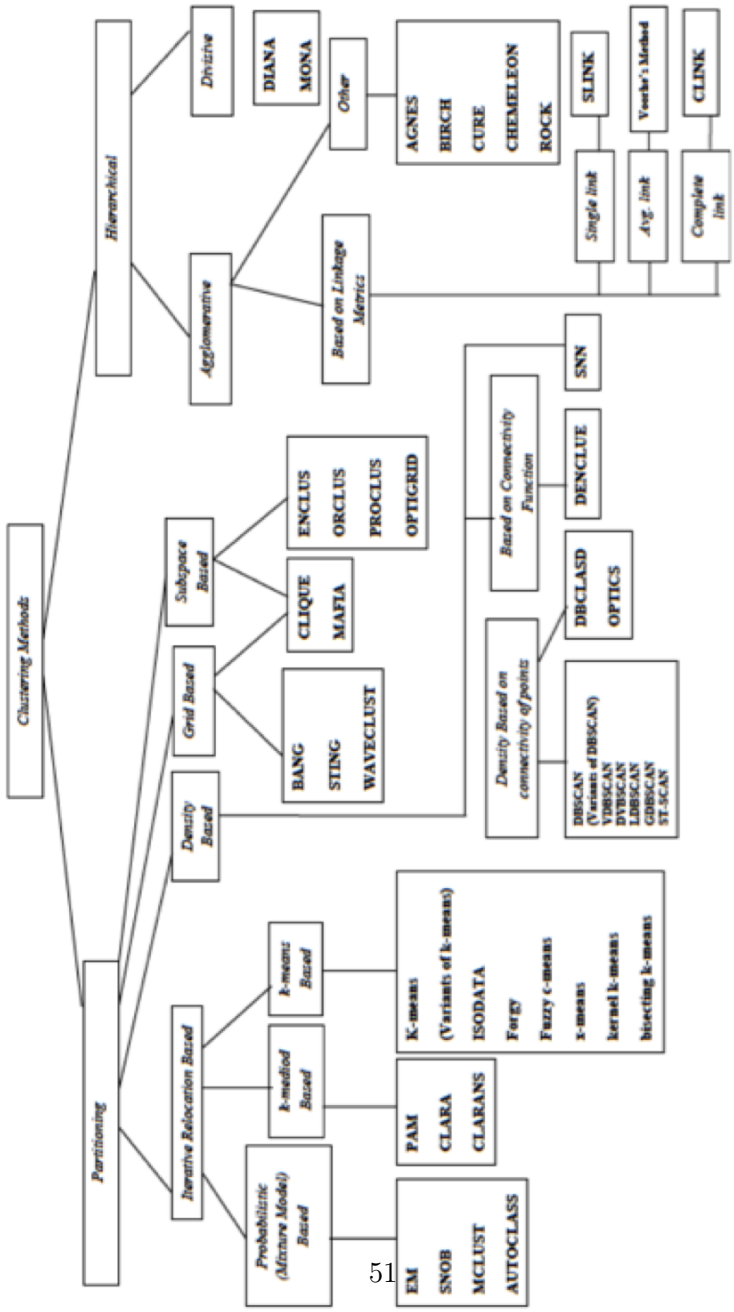


Figura 2.2: Classificazione dei metodi di *Clustering*

quanto il lessico nell'intera collezione può essere nell'ordine di 10^5, ma un singolo documento può contenere solo poche centinaia di parole. Se i documenti da raggruppare sono molto brevi, come nel caso di frasi o di *tweet*, questo problema si presenta in maniera ancor più incisiva.
Un altro aspetto da cui non si può prescindere è che il numero di termini nei differenti documenti può variare considerevolmente, ciò implica che nel processo di *Clustering* bisogna tener conto della diversa lunghezza dei documenti normalizzando opportunamente la loro rappresentazione.
Inoltre i termini all'interno della raccolta sono tipicamente correlati tra loro, pertanto il numero di concetti distinti nei dati è molto più piccolo del numero di termini effettivamente presenti. In tal senso è necessario far riferimento ad appropriate tecniche di *feature selection* o di *feature trasformation* o ancora a metodi specifici di *Clustering* che tengano conto di questa correlazione.
Nell'ambito di questo lavoro faremo riferimento unicamente a metodi di *Clustering* applicabili su documenti codificati secondo lo schema BOW (*Bag Of Words*), ma è bene tener presente che i documenti possono essere codificati anche come stringhe di testo, pertanto con una diversa rappresentazione è necessario implementare specifiche classi di algoritmi di *Clustering*.
Seguendo quindi una classificazione analoga a quella dei metodi di *Clustering* classici ed escludendo metodi non adatti per questo specifico *task*, potremmo comunque suddividere i metodi per il *Clustering* di documenti in metodi gerarchici e metodi partitivi.

2.4.1 Metodi gerarchici

I metodi gerarchici sono stati ampiamente studiati nella letteratura del *Clustering* [42]. I metodi gerarchici producono una gerarchia di partizioni an-

nidate, dove ogni livello intermedio può essere visto come la combinazione di due *cluster* del successivo livello più basso (o come la suddivisione di un *cluster* di livello più alto).
Il risultato di un algoritmo di *Clustering* gerarchico può essere rappresentato graficamente attraverso un diagramma ad albero, chiamato *dendrogramma*. Questo grafico consente di visualizzare il processo di fusione e le partizioni ottenute nei livelli intermedi. Nell'ambito del *Clustering* di documenti, il *dendrogramma* fornisce una tassonomia, o un indicizzazione gerarchica.
In prima istanza questi metodi si distinguono rispetto alla strategia secondo cui si concretizza il processo di *Clustering* in *agglomerativi* (o ascendenti) e *divisivi* (o discendenti).
Secondo il primo approccio le unità sono raggruppate successivamente in *cluster* partendo dall'insieme iniziale dei dati, per cui ogni unità appartiene ad un *cluster* distinto, fino ad arrivare ad un unico *cluster* in cui confluiscono tutte le unità del collettivo considerato.
Al contrario nel caso di un approccio discendente, tutte le unità del collettivo appartengono inizialmente ad un unico *cluster*, per poi essere suddivise in *cluster* sempre più piccoli fino alla situazione diametralmente opposta in cui ogni unità appartiene ad un *cluster*.
Tra i due approcci, i metodi gerarchici agglomerativi più tradizionali, sebbene rappresentino delle proposte generali sono stati ritenuti particolarmente idonei per il *Clustering* di documenti [84] [88].
Il concetto generale del *Clustering* gerarchico agglomerativo è, quindi, raggruppare, ad ogni passo dell'algoritmo, i documenti in *cluster* in base alla loro somiglianza reciproca.
Il processo di aggregazione dei documenti produce di conseguenza una struttura gerarchica che fornisce un'immagine delle relazioni fra i documenti e

dove è possibile valutare i livelli di similitudine tra i diversi *cluster* cui gli stessi documenti appartengono.
I documenti molto simili sono infatti raggruppati in piccoli *cluster* ai livelli più bassi, e documenti più basicamente collegati in *cluster* più grandi e generici ai livelli più alti, fino ad arrivare al nodo dell'albero in cui confluiscono tutti i documenti della raccolta. Tutti i metodi gerarchici agglomerativi infatti fondono i gruppi in successione tenendo conto della "migliore" similarità (distanza) a coppie tra gruppi di documenti. In generale un algoritmo gerarchico agglomerativo opera iterativamente secondo il seguente schema:

1. viene calcolata la matrice di similarità, in cui ogni cella contiene il valore della similarità tra le diverse coppie di documenti;

2. i due *cluster* per cui si rileva la maggiore similarità sono raggruppati in un unico *cluster*;

3. si aggiorna la matrice di similarità allo scopo di calcolare la similarità a coppie tra il nuovo *cluster* e i gruppi originari;

4. si ripetono gli step 2 e 3 finchè tutti i documenti appartengono ad un solo *cluster*.

La principale differenza tra i diversi algoritmi che discendono da questo concetto generale sta nel modo in cui questa similarità a coppie è calcolata tra gruppi di documenti.
I metodi per la fusione dei gruppi maggiormente utilizzati nell'ambito del *Text Clustering*, ai fini delle aggregazioni successive sono il *Single linkage Clustering*[84][16], il *Complete linkage Clustering* e l'*Average linkage Clustering*. Di seguito sono descritte le peculiarità di ciascuno degli approcci.

Single linkage Clustering Con questo tipo di approccio la similarità tra due gruppi di documenti C_i e C_j è definita come la più grande similarità tra tutte quelle calcolabili rispettivamente tra ciascun documento del gruppo i e ciascun documento del gruppo j. Questo modo di procedere è equivalente al metodo del legame singolo, formalmente, quindi, la similarità tra due gruppi diversi è calcolata:

$$S(C_i, C_j) = \max_{x \in C_i, y \in C_j} s(x, y) \tag{2.10}$$

dove x e y sono due documenti appartenenti rispettivamente al cluster C_i e al cluster C_j. Il *Single linkage Clustering* ha il vantaggio di essere in pratica estremamente efficiente, poiché dopo aver calcolato tutte le similarità tra coppie di documenti, riordinandole in senso decrescente, queste sono processate dall'algoritmo in un ordine prestabilito, evitando così di inserire nel calcolo le similarità nulle. Sebbene sia un metodo particolarmente efficace per la ricerca di gruppi di forma allungata, un possibile effetto collaterale di questo metodo è il *concatenamento* tra documenti che in realtà appartengono a gruppi diversi. Ciò può condurre all'individuazione di *cluster* poco significativi, specialmente ai livelli più alti della gerarchia.

Complete linkage Clustering In questa tecnica la similarità tra due gruppi di documenti C_i e C_j è definita come la minima similarità calcolata tra ogni coppia di documenti appartenenti ai due diversi gruppi. Formalmente è possibile esprimerla come:

$$S(C_i, C_j) = \min_{x \in C_i, y \in C_j} s(x, y) \tag{2.11}$$

dove x e y sono due documenti appartenenti rispettivamente al cluster C_i e al cluster C_j. Al contrario del *Single linkage Clustering*

con questo approccio per il calcolo della similarità tra gruppi si evita il concatenamento di gruppi distinti, poiché due documenti molto diversi tra loro non verranno inclusi nello stesso*cluster*. Tuttavia il *Complete linkage Clustering* è computazionalmente più oneroso del precedente, in quanto richiede $O(n^2)$ in termini di spazio e $O(n^3)$ in termini di tempo, dove n è il numero totale di nodi. Questo problema nel contesto del *Clustering* di documenti è senz'altro meno pressante poiché molte delle similarità tra coppie saranno nulle.

Average linkage Clustering Questo approccio rappresenta un compromesso tra i due algoritmi precedenti, in tal caso, infatti, la similarità tra due gruppi di documenti C_i e C_j è definita come media delle similarità tra le coppie di documenti appartenenti ai due diversi gruppi:

$$S(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i, y \in C_j} s(x, y) \tag{2.12}$$

dove x e y sono due documenti appartenenti rispettivamente al cluster C_i e al cluster C_j. Chiaramente questo modo di procedere è meno efficiente in termini di tempo rispetto al *Single linkage Clustering*, in quanto è necessario calcolare la similarità tra un gran numero di coppie di documenti per determinarne una media. Al contempo risulta però più robusto in termini di qualità del *Clustering*, evitando inoltre il problema del concatenamento. É possibile ottenere una riduzione dei tempi di calcolo nell'*Average linkage Clustering* approssimando la similarità media tra due *cluster* C_i e C_j, calcolando la similarità tra il "documento-medio" di C_i e il "documento-medio" di C_j. Formalmente

questo concetto è esprimibile come segue:

$$S(C_i, C_j) = s(\bar{c}_i, \bar{c}_j) \tag{2.13}$$

dove $\bar{c}_i$ e $\bar{c}_j$ sono i vettori-documento medi nei due gruppi. Questo approccio non funziona egualmente bene per tutti i tipi di dati, ma è particolarmente efficace nel caso dei dati testuali, poiché il centroide del *cluster* (documento-medio) è semplicemente il concatenamento dei documenti in esso inclusi.

Il vantaggio di questi metodi è non solo la possibilità di visualizzare graficamente le relazioni tra i documenti a diversi livelli di dettaglio, ma anche il non dover decidere *a priori* il numero di *cluster* in cui raggruppare i documenti, sebbene siano purtroppo poco scalabili, quindi, generalmente meno adatti per *corpora* testuali di grandi dimensioni. Al contempo la caratteristica che accomuna i diversi algoritmi gerarchici è che una volta che due gruppi di documenti sono stati aggregati non saranno più separati in seguito. Inoltre, applicando uno qualunque di questi algoritmi, si fa riferimento per ogni valore del numero di gruppi k sempre alla stessa struttura di raggruppamento (dendrogramma), guardando unicamente ad un diverso livello di similarità nella struttura stessa, constatazione che evidenzia una struttura molto più rigida rispetto a quella ottenibile con metodi partitivi. In aggiunta è importante sottolineare l'opinione ampiamente condivisa secondo cui i metodi partitivi diano risultati migliori in termini di qualità del *Clustering* [78].

2.4.2 Metodi partitivi

Al contrario dei metodi gerarchici, i metodi di *Clustering* partitivi creano un'unica partizione dei dati, dove il numero k di gruppi è uno dei parametri

dell'algoritmo, che va solitamente specificato dall'operatore.
I metodi di *Clustering* partitivi possono individuare una partizione in k *cluster* di un insieme di oggetti in maniera "diretta" oppure tramite una sequenza di ripetute suddivisioni.
In generale, un algoritmo partitivo che produce direttamente una partizione in k gruppi opera nel modo seguente. Si seleziona un insieme di k oggetti dal collettivo iniziale come rappresentativi dei k *cluster* da individuare, definiti semi di partenza. Si calcola poi la similarità tra ogni oggetto del collettivo e ciascuno dei k semi e si assegna ciascun oggetto al *cluster* corrispondente al seme più vicino in termini di similarità. Questa procedura consente di individuare un'iniziale partizione in k *cluster*, che sarà rifinita attraverso ripetute iterazioni finchè la funzione criterio scelta converga, nella maggior parte dei casi, ad un'ottimo locale.
Tra gli algoritmi partitivi "diretti" che assegnano iterativamente gli oggetti ai k *cluster* il *Kmeans* [58] ne rappresenta un classico esempio. Inoltre, sebbene si è più volte sottolineato quanti nuovi algoritmi sono stati pubblicati in oltre cinquant'anni di ricerca scientifica, il *Kmeans* è ancora un importante riferimento nella letteratura del *Clustering* [44] e in particolare ampiamente utilizzato per il *Clustering* di documenti.
Secondo questo metodo gli oggetti, rappresentati in uno spazio *p-dimensiona le*, sono suddivisi in k *cluster*, in modo tale da minimizzare un certo criterio di convergenza. Tipicamente si minimizza l'errore quadratico, ossia la somma della distanza di ciascun oggetto dal centro del proprio *cluster*, chiamato centroide. É bene precisare che il centroide di un*cluster* è sostanzialmente una media pesata degli oggetti che vi appartengono. Nella maggior parte dei casi questo non corrisponde ad alcun oggetto del dataset d'origine. Il modo in cui calcolare il centro del *cluster* è ciò che differenzia il *Kmeans* da

metodi partitivi *Kmedoids* [46], dove i centroidi sono oggetti dell'insieme di partenza, cioè sono gli oggetti che hanno la minima distanza media dagli altri oggetti appartenenti al loro stesso *cluster*. Sebbene tali metodi siano più robusti rispetto alla presenza di *outliers*, risultano meno idonei per la categorizzazione di documenti in quanto una gran parte di coppie di documenti non hanno molti termini in comune, e di conseguenza le rispettive similarità tra queste coppie sono molto piccole. Pertanto, un unico documento/medoide spesso non contiene tutti i concetti necessari per costruire effettivamente un *cluster* intorno ad esso.
L'algoritmo *Kmeans* opera iterativamente secondo il seguente schema:

1. si scelgono arbitrariamente k centroidi come soluzione iniziale;
2. si calcolano del *cluster* di appartenenza di ciascun documento in accordo con l'attuale scelta dei k centroidi;
3. si calcolano i nuovi centroidi di ogni *cluster*;
4. si ripetono gli step 2 e 3 finchè non sia soddisfatto il criterio di arresto.

Il raggiungimento dell'ottimo globale, ossia dei k centri migliori, è un problema NP completo, quindi il criterio di convergenza prevede di arrestarsi dopo un certo numero di iterazioni, o qualora il miglioramento della soluzione non è più apprezzabile. Il *Kmeans* ha una complessità computazionale $O(nkl)$, dove n è il numero di documenti nel *corpus*, k il numero di gruppi desiderato e l è il numero di iterazioni. Il principale vantaggio di questo metodo è la capacità di gestire dataset di notevoli dimensioni, raggiungendo la convergenza attraverso un estremamente piccolo numero di iterazioni. Tuttavia il metodo è sensibile alla presenza di *outliers* e di rumore nei dati e risente della scelta dei centroidi iniziali. Inoltre, il centroide di un dato

cluster di documenti può contenere un gran numero di termini, questo rende di certo rallenta il calcolo delle similarità nella successiva iterazione.
Per far fronte a tali problemi sono state sviluppate diverse varianti del *Kmeans*, alcune delle quali specificatamente indirizzate al trattamento dei dati testuali [40][86]. Tra le diverse proposte particolarmente interessante risulta quella di Dhillon e Modha [20] che, per ridurre l'effetto derivante dalla differente lunghezza dei documenti, suggeriscono di utilizzare come misura di similarità il *coseno* (vedi paragrafo 2.3.1). Il metodo proposto, noto in letteratura come *Spherical Kmeans*, è anch'esso una variante del classico *Kmeans*, in cui la distanza Euclidea è sostituita dalla seguente espressione:

$$d(D_1, \bar{D}) = 1 - s_C(D_1, \bar{D}) = 1 - \frac{\left|D_1 \cap \bar{D}\right|}{\sqrt{|D_1| \times \left|\bar{D}\right|}} \tag{2.14}$$

dove $\bar{D}$ è il centroide del *cluster* cui appartiene D_1. La normalizzazione dei vettori documento implica che gli stessi possano essere visti come punti in una sfera altamente dimensionale, pertanto il metodo ne individuerà una partizione, sezionando tale sfera mediante una insieme di *hypercircle*. Per ognuno dei *cluster* finali si ottiene un centroide di norma unitaria (*concept vector*) che riassume l'informazione semantica contenuta nel gruppo che rappresenta. Lo *Spherical Kmeans* ha un certo numero di vantaggi dal punto di vista computazionale sia per la quanto riguarda la sparsità che caratterizza i dati testuali, sia perché può essere facilmente parallelizzato, convergendo ad un ottimo locale in un numero esiguo di iterazioni. Inoltre, dal punto di vista statistico il metodo genera dei *concept vector* che possono essere utilizzati per classificare successivamente nuovi documenti.

Un algoritmo partitivo che opera, invece, tramite delle successive suddivisio-

ni del collettivo preso in considerazione individua la soluzione di *Clustering* in k gruppi nel modo seguente. Inizialmente gli oggetti sono suddivisi in 2 *cluster*, applicando un algoritmo partitivo "diretto"; si seleziona uno dei due *cluster* ottenuti e si ripete la suddivisione. Il processo di selezione di uno dei *cluster* e di successiva suddivisione dello stesso si ripete $k-1$ volte, individuando in tal modo la partizione in k *cluster*. Ognuna delle bisezioni è effettuata affinché la suddivisione risultante in 2 gruppi ottimizzi una particolare funzione criterio.
La differenza sostanziale rispetto ad un algoritmo di *Clustering* partitivo diretto è che in tal caso il passaggio chiave è il criterio utilizzato per selezionare ad ogni *step* del metodo il *cluster* da suddividere.
Inoltre, un approccio di bisezioni successive, similmente a quanto accade per i metodi gerarchici, produce come risultato una gerarchia di partizioni annidate, che con un metodo partitivo "diretto" è ottenibile solo rieseguendo l'algoritmo $k-1$ volte, al variare del numero di gruppi k.
Un algoritmo che opera seguendo questo criterio, rivelatosi particolarmente efficace per il *Clustering* di documenti [78], è il *bisecting Kmeans*.
Il *bisecting Kmeans* è solitamente etichettato come una variante del *Kmeans* e convenzionalmente appartiene alla famiglia degli algoritmi partitivi, sebbene possa essere visto sostanzialmente di un algoritmo gerarchico divisivo.
Il *bisecting Kmeans* ha una complessità computazionale $O(n)$, dove n è il numero di documenti nella collezione.
Inoltre, se il numero di *cluster* è grande e non si effettua alcun *"raffinamento"* della soluzione di *Clustering*, questo algoritmo risulta più efficiente del *Kmeans* classico, poiché in tal caso non è necessario valutare la similarità di ogni documento rispetto ad ognuno dei k centroidi in quanto per suddividere un *cluster*, basta tener conto dei documenti inclusi nel *cluster* e delle

loro similarità rispetto a soli 2 centroidi. Calcolare il prodotto scalare tra un documento e il centroide del *cluster* è infatti equivalente a calcolare la similarità media tra quello stesso documento e tutti i documenti inclusi nel *cluster* che quel centroide rappresenta.

Capitolo 3

Tecniche di validazione dei metodi di Clustering

La *Cluster Analysis*, facendo parte delle tecniche di analisi multidimensionale di tipo esplorativo, non necessita di alcun tipo di assunzione *a priori*. Si suppone l'esistenza di una struttura di gruppo, e l'obiettivo è quindi individuare la naturale classificazione dei dati. D'altra parte, prima, durante e dopo l'analisi, il ricercatore deve prendere alcune decisioni, quali ad esempio la scelta delle variabili, del criterio di similarità, dell'algoritmo, del numero di *cluster* da ottenere. Trattandosi, come detto, di una tecnica di tipo esplorativo, la valutazione della qualità della soluzione è a volte considerata una questione estranea a questo tipo di approccio. In alcuni casi, le tecniche di *Clustering* producono risultati inadeguati, individuando dei raggruppamenti che non riflettono la struttura naturale dei dati.
Una delle motivazioni principali risiede nel fatto che la maggior parte degli algoritmi di *Clustering* individua una partizione anche se i dati si distribuiscono casualmente e non è lecito pensare ad un eventuale raggruppamento.

Inoltre, gli algoritmi agiscono in maniera diversa a seconda della natura dei dati, delle ipotesi iniziali per definire gruppi e dei parametri di input, valutazioni che spesso gli utenti meno esperti trascurano. Vari studi proposti in letteratura [64] dimostrano che strategie di *Clustering* differenti conducono spesso a risultati non dissimili, eventualità che potrebbe riflettere la presenza di una "forte" struttura di gruppo. Al contempo in molte applicazioni pratiche i risultati sembrano dipendere, non solo dal metodo di raggruppamento usato, ma anche dalle trasformazioni applicate ai dati: la stessa standardizzazione può far si che la partizione ottenuta sia sensibilmente diversa da quella ottenuta a partire dai dati d'origine.
Pertanto se i dati presi in esame soddisfano le assunzioni dell'algoritmo (il più delle volte non esplicitate) riguardo a ciò che s'intende per cluster (naturali), quest'ultimo riuscirà senza alcun problema ad individuare una struttura sottostante i dati. Il problema è sostanzialmente quello di decidere se la soluzione ottenuta riflette la struttura naturale dei dati o è stata indotta dall'algoritmo di *Clustering* scelto. D'altra parte l'assenza di gruppi noti *a priori* in un processo di *Clustering* ha reso difficile trovare un indicatore adeguato per valutare se la soluzione ottenuta, il numero di cluster, la loro forma è ammissibile o meno.
Per dare una risposta a domande come: "Esiste una struttura non casuale nei dati?", "Quanti gruppi ci sono nel set di dati?", "La partizione ottenuta è conforme alla struttura dei nostri dati?", "C'è una ripartizione migliore per il nostro dataset?" è necessaria allora una fase di valutazione dei risultati ottenuti. L'insieme delle tecniche che mirano ad una valutazione quantitativa e oggettiva dei risultati di un processo di *Clustering* va sotto il nome di *Cluster validity methods* [35].
In questo capitolo saranno descritti i principali approcci per la validazione

di una soluzione di *Clustering*.

3.1 Il concetto di validazione

In Statistica, validare il risultato di un metodo vuol dire valutare in maniera quantitativa e oggettiva se ciò che è stato individuato nella ricerca rispecchia effettivamente il fenomeno indagato. In tal senso un risultato è valido se rappresenta la migliore approssimazione possibile della realtà [15].
Nell'ambito del *Clustering* la validazione di una partizione o di una gerarchia di partizioni annidate può essere effettuata tenendo conto dei seguenti criteri [74]:

- oggettività, per cui i ricercatori che lavorano indipendentemente sullo stesso insieme di dati devono giungere agli stessi risultati;
- stabilità dei risultati del *Clustering*, operando su dati equivalenti;
- capacità predittiva delle variabili su un nuovo insieme di dati.

Jain e Dubes ritengono valida una soluzione di *Clustering* nella misura in cui la partizione o la gerarchia individuata fornisce una vera informazione sui dati o, in altri termini quanto tale soluzione sia in grado di riflettere la caratteristiche intrinseche dei dati [42]. Halkidi *et al.* definiscono lo schema di raggruppamento ottimale come il risultato di un algoritmo di *Clustering* che meglio si adatta alla struttura sottostante i dati [35]. In accordo con tali definizioni i metodi di *Cluster Validation* mirano a valutare la bontà di un risultato di *Clustering* cercando di stabilire quanto lo schema di raggruppamento ottenuto rifletta la naturale struttura dei dati. In particolare, un processo di validazione può essere indirizzato a:

- determinare se esiste una struttura non casuale nei dati;
- determinare il numero corretto di *cluster* presenti nei dati;
- valutare quanto i risultati ottenuti da una *Cluster Analysis* si adattano ai dati senza avere a disposizione informazioni esterne;
- confrontare i risultati di una *Cluster Analysis* con informazioni note *a priori*, quali, ad esempio, etichette di classe fornite esternamente;
- confrontare differenti soluzioni per determinare quale sia la migliore.

Per ciascuno di questi scopi sono state sviluppate una serie di proposte descritte in dettaglio nei seguenti paragrafi.

3.2 Le misure per la validazione del *Clustering*

In letteratura sono tradizionalmente individuati tre approcci per investigare la validità del *Clustering*:

- criteri di validazione **esterni**, che misurano quanto i *cluster* individuati corrispondono a etichette di classe fornite esternamente (conoscenza pregressa);
- criteri di validazione **interni**, che misurano quanto una soluzione di *Clustering* si adatta bene ai dati, quando i dati sono la sola informazione disponibile;
- criteri di validazione **relativi**, che misurano la bontà di una soluzione di *Clustering* confrontandola con i risultati ottenuti da altri algoritmi di *Clustering* o dallo stesso algoritmo, ma usando differenti valori dei parametri.

Sulla base di questi tre criteri, che definiscono delle strategie generali, sono stati definiti una serie di indici di validità [35], anche se, spesso, nell'ambito della *cluster validity* le misure utilizzate sono considerate criteri piuttosto che indici. Pertanto, si può indicare il criterio come la strategia generale, mentre l'indice è la misura numerica che la implementa.
Per il calcolo degli indici di validazione esterna sono necessarie le informazioni circa le etichette di classe degli oggetti su cui si esegue il *Clustering*. Tali indici permettono di misurare la corrispondenza tra l'etichetta calcolata del *cluster* e l'etichetta della classe, nota *a priori*. Gli indici che rientrano in questa categoria, possono essere classificati in indici orientati alla classificazione e indici basati sulla similarità. I primi sono indicatori nati nell'ambito della classificazione supervisionata e adattati agli scopi del *Clustering*, gli indici basati sulla similarità si fondano sulla premessa che se due oggetti sono nello stesso *cluster* allora devono appartenere alla stessa classe e *vice versa*. Quest'ultimo approccio alla validazione dei *cluster* può essere visto in termini di confronto tra due matrici: una matrice di similarità tra i *cluster* e una matrice di similarità tra le classi, definita a partire dalle etichette di classe, in cui il generico elemento ij delle matrici avrà valore 1 se gli oggetti, i e j, appartengono rispettivamente allo stesso *cluster* o alla stessa classe e 0 altrimenti. Gli indici di validazione interna si basano sui concetti di coesione (Cluster Cohesion) e di separazione (Cluster Separation). Questi indicatori, infatti, misurano quantitativamente quanto la partizione ottenuta risponde all'obiettivo del *Clustering*, individuazione di gruppi coesi e ben separati tra loro.
Alcuni degli indici di validazione sia esterna che interna [83] sono spesso usati per la validazione mediante criteri relativi. In questo caso, infatti, si valutano i risultati dei differenti schemi di *Clustering* al variare dei parame-

tri, attraverso un opportuno indice di validazione, identificando la migliore partizione dei dati. In generale un indice di validazione relativo è una misura di validazione supervisionata o non supervisionata, usata per eseguire un confronto. Pertanto gli indici relativi non sono realmente un gruppo separato di misure di validazione, ma rappresentano uno specifico uso degli indici interni ed esterni.

3.2.1 Misure esterne per la validazione

L' approccio esterno alla validazione presuppone la conoscenza del *vero* numero di gruppi e della loro composizione. Se si hanno a disposizione le etichette di classe, si esegue il *Clustering* per comparare i risultati provenienti dall'applicazione di diversi algoritmi, con l'obiettivo di individuare l'algoritmo ottimale per uno specifico dataset. Inoltre, gli indici che rientrano in questa categoria sono particolarmente utilizzati nell'ambito della classificazione supervisionata, allo scopo di valutare la performance del singolo metodo di classificazione, in termini di accuratezza della classificazione ottenuta, ossia quanto i *cluster* individuati corrispondono a etichette di classe fornite esternamente.
Le misure esterne per la validazione, possono essere classificate secondo due approcci:

- **Classification-oriented**, che valutano in che misura i *cluster* individuati contengono oggetti appartenenti alla stessa classe;

- **Similarity-oriented**, che misurano con quale frequenza due oggetti che appartengono allo stesso *cluster*, appartengono effettivamente alla stessa classe.

Nei seguenti due paragrafi sono descritti i principali indici che afferiscono a ciascuna delle precedenti categorie.

Gli indici esterni *classification-oriented* Nell'ambito della classificazione supervisionata sono stati proposti un gran numero di indici per valutare la performance di un classificatore. In tal caso si misura il grado di concordanza tra le etichette di classe predette e la classificazione reale dei dati. Tuttavia se si utilizza una soluzione di *Clustering*, al posto delle etichette di classe predette attraverso un metodo di classificazione supervisionata, l'uso che si fa di questi indicatori sostanzialmente non cambia. Gli indici *classification-oriented* più utilizzati per la validazione esterna dei *cluster* sono l'F-measure [53], l'Entropia e la Purezza [72].
L'**F-measure** combina due misure comunemente utilizzate per la valutazione di un sistema di *Information Retrieval*: la *precision* e la *recall*.
La *precision* è la proporzione di oggetti nel *cluster* i che appartengono ad una specifica classe. La *precision* del *cluster* i rispetto ad una generica classe j è:

$$precision(i, j) = p_{ij} = \frac{n_{ij}}{n_i} \tag{3.1}$$

La *recall* valuta in che misura un *cluster* contiene oggetti di una specifica classe ed è calcolata come rapporto tra il numero di oggetti nel *cluster* i che appartengono alla classe j e il numero di oggetti nella classe j:

$$recall(i, j) = \frac{n_{ij}}{n_j} \tag{3.2}$$

L'*F-measure* è una combinazione dei due indicatori precedenti, nello specifico si tratta della loro media armonica, e può essere calcolata per ciascuno dei *cluster* in rapporto ad ognuna delle classi.

In generale, l'*F-measure* del *cluster* i rispetto alla classe j è:

$$F(i,j) = \frac{2 \times precision(i,j) \times recall(i,j)}{precision(i,j) + recall(i,j)} \tag{3.3}$$

I valori assunti dall'indice variano tra 0 e 1, più sono prossimi ad 1 migliore risulta la corrispondenza tra i risultati del *Clustering* e la classificazione *a priori* e di conseguenza maggiore sarà la qualità della soluzione.

L'**Entropia** misura la purezza delle etichette all'interno dei *cluster*. Pertanto, se tutti i cluster consistono di oggetti con una sola etichetta, l'entropia è zero. In ogni caso, come variano le etichette nel cluster, così aumenta l'entropia. Dopo aver calcolato la distribuzione delle etichette di classe, l'entropia di ciascun *cluster* è definita come segue:

$$E_i = \sum_j p_{ij} \log(p_{ij}) \tag{3.4}$$

dove p_{ij} è la *precision* del *cluster* i rispetto alla classe j, ossia la proporzione di oggetti della classe j nel *cluster* i.

Come mostrato nella seguente equazione l'entropia totale del dataset è la somma pesata delle entropie di tutti i *cluster*:

$$E = \sum_{i=1}^{k} \frac{n_i}{n} E_i \tag{3.5}$$

dove k è il numero dei *cluster* e n_i è la dimensione del *cluster* i.

Il concetto di **Purezza** è molto simile a quello di entropia, in quanto anche questa misura ha l'obiettivo di valutare quantitativamente il grado in cui ciascun *cluster* contiene oggetti appartenenti ad una sola classe. Facendo riferimento alla precedente terminologia, la purezza del *cluster* i è calcolata come:

$$P_i = \max_j p_{ij} \tag{3.6}$$

La purezza complessiva di una soluzione di *Clustering* è calcolata come somma pesata della purezza di ciascuno dei *cluster*:

$$P = \sum_{i=1}^{k} \frac{n_i}{n} P_i \tag{3.7}$$

dove, anche in tal caso, i pesi sono espressi come percentuale di oggetti del dataset in ciascun gruppo.

Gli indici esterni *similarity-oriented* Secondo l'approccio *similarity-oriented*, gli indici esterni forniscono una misura dell'accuratezza dei risultati, in termini di quante osservazioni sono correttamente classificate secondo le etichette fornite *a priori* e quante, invece risultano appartenenti ad una classe cui non dovrebbero essere associate. I principali indicatori basati sulla similarità sono l'indice di Jaccard [41], l'indice di Rand [64] e l'indice Fowlkes-Mallows [29].
L'indice di **Jaccard** misura il livello di accordo tra un insieme di etichette di classe C e una generica partizione K, ottenuta a seguito di una *Cluster analysis*, determinando il numero di coppie di punti assegnati allo stesso *cluster* in entrambe le partizioni:

$$J(C, K) = \frac{a}{a + b + c} \tag{3.8}$$

dove a è il numero di coppie di punti con la stessa etichetta in C che sono assegnati allo stesso *cluster* in K, b è il numero di coppie di punti con la stessa etichetta, ma classificate in differenti *cluster* e c, invece, è il numero di coppie di punti nello stesso *cluster*, ma che presentano etichette di classe differenti. L'indice varia tra 0 e 1, pertanto il massimo valore si ottiene quando C e K sono identici.

Come per l'indice di Jaccard, l'indice di **Rand**, misura il grado di accordo tra le etichette di classe fornite *a priori* e una generica partizione K. Tuttavia può essere utilizzato anche per la comparazione di due schemi di *Clustering*, qualora non fossero note le etichette delle classi. La differenza che intercorre tra i due indicatori è che l'indice Rand, tiene conto non solo del numero di coppie di punti assegnati allo stesso *cluster* in entrambe le partizioni, ma anche del numero di coppie di punti con differenti etichette di classe che sono assegnate a *cluster* differenti. Formalmente:

$$R(C, K) = \frac{a + d}{a + b + c + d} \tag{3.9}$$

dove d è il numero di coppie di punti con diversa etichetta in C che sono assegnati a differenti *cluster* in K. L'indice varia tra 0 e 1, più alti sono i valori che assume più c'è similarità tra le due partizioni. Il massimo dell'accuratezza si ottiene quando l'indice è prossimo ad 1. Il problema dell'indice di Rand è che il suo valore atteso, per confronti tra partizioni casuali non è costante (ad esempio zero), pertanto Hubert e Arabie (1985) [39] propongono una versione modificata dell'indicatore, l'**adjusted-Rand** che supera questo inconveniente.
Un altro indice esterno *similarity-oriented* è l'indice di **Fowlkes-Mallows** [29]. Questa misura di similarità può essere utilizzata sia per la comparazione di due schemi di *Clustering* gerarchico, sia per comparare una classificazione C nota *a priori* con una partizione K. Con riferimento alle misure esterne fin qui presentate l'indice Fowlkes-Mallows rappresenta sostanzialmente la media geometrica di *precision* e *recall*.

3.2.2 Misure interne per la validazione

Nella maggior parte degli scenari applicativi, non sono disponibili informazioni riguardanti il *vero* numero di *cluster*, o addirittura la composizione dei gruppi, in tal caso gli indici di validazione interna rappresentano l'unica opzione disponibile per valutare la qualità di una soluzione di *Clustering*. Queste misure sono utilizzate principalmente per scegliere, fissato il criterio di raggruppamento, la struttura di raggruppamento più idonea per uno specifico set di dati.
In generale la validità complessiva di una soluzione di *Clustering* è considerata una combinazione lineare della validità di ciascun cluster.
Formalmente:

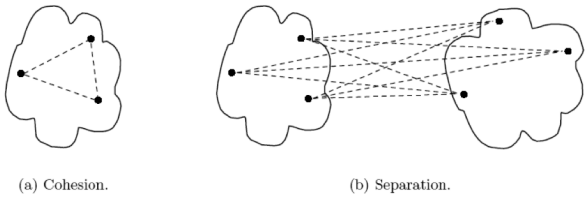
$$overall\ validity = \sum_{j=1}^{k} w_j validity(C_j) \tag{3.10}$$

Poiché l'obiettivo della *Cluster Analysis* è l'individuazione di gruppi coesi e ben separati, gli indici di validazione interna si basano principalmente sui seguenti due criteri [79][91]:

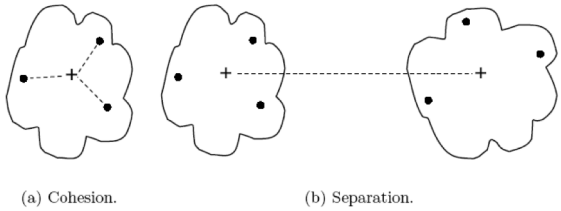
- **Cluster Cohesion**, che misura l'affinità tra gli oggetti di un cluster;
- **Cluster Separation**, che misura quanto i cluster sono distinti e ben separati rispetto agli altri cluster.

Pertanto la validità complessiva di una soluzione sarà funzione della coesione, della separazione o di entrambe. I pesi w_j attribuiti a ciascun *cluster* dipenderanno dall'indice di validazione scelto. In alcuni casi i *cluster* assumeranno tutti lo stesso peso, in altri il peso sarà attribuito in relazione alla numerosità di ciascun *cluster*, in altri ancora rifletteranno delle proprietà più complesse. In linea generale se l'indice misura la coesione dei *cluster*,

valori alti saranno migliori, se al contrario l'indice misura la separazione saranno auspicabili valori bassi.
La coesione di ciascun *cluster* e la separazione tra i *cluster* possono essere calcolate sia per rappresentazioni basate su grafi sia per rappresentazioni basate su prototipi. Pertanto gli indici di validazione interna possono essere classificati in misure *graph-based* e *prototype-based*. Nel primo caso, (Fig.3.1(a)) la coesione di un *cluster* può essere definita come la somma dei pesi degli archi nel *proximity graph* che connettono gli oggetti entro il *cluster*. In un *proximity graph*, infatti, ciascun oggetto è considerato un



(a) Graph-based



(b) Prototype-based

Figura 3.1: Rappresentazione grafica dei concetti di coesione e separazione

nodo, ogni coppia di oggetti è connessa tramite un arco e ad ogni arco è assegnato un peso, pari alla prossimità tra i due oggetti. Allo stesso modo la separazione può essere misurata dalla somma dei pesi degli archi tra nodi appartenenti a *cluster* distinti.

Matematicamente, per rappresentazioni basate su grafi, i due concetti sono espressi rispettivamente nelle equazioni 3.11, 3.12.

$$cohesion(C_j) = \sum_{x,y \in C_j} proximity(x, y) \tag{3.11}$$

$$separation(C_i, C_j) = \sum_{x \in C_i y \in C_j} proximity(x, y) \tag{3.12}$$

Per quanto riguarda le misure *prototype-based* (Fig.3.1(b)), la coesione di un *cluster* può essere definita come la somma delle prossimità di tutti i punti entro il *cluster* rispetto ad al prototipo (centroide o medoide) del *cluster*. In maniera analoga la separazione tra due *cluster* può essere misurata in termini di prossimità tra i loro due prototipi. Matematicamente, per rappresentazioni basate su prototipi, la coesione è espressa nell'equazione 3.13, mentre la separazione si può calcolare in due modi, poiché la separazione dei prototipi dei *cluster* dal prototipo (centro) dell'intero dataset 3.14 è talvolta direttamente legata alla separazione dei cluster tra loro 3.15.

$$cohesion(C_j) = \sum_{x \in C_j} proximity(x, c_j) \tag{3.13}$$

$$separation(C_i, C_j) = proximity(c_i, c_j) \tag{3.14}$$

$$separation(C_i) = proximity(c_i, c) \tag{3.15}$$

Per entrambi gli approcci la funzione di prossimità può essere una similarità, una dissimilarità o una funzione di entrambe. É bene precisare che in letteratura [80] [51] oltre alle misure basate su coesione e separazione sono stati proposti anche altri indici di validazione interna, tuttavia alcuni di questi hanno dimostrato di non essere particolarmente affidabili, mentre

altri sono stati realizzati per trattare dati caratterizzati da una particolare struttura. Nei seguenti paragrafi saranno presentati i principali indici di validazione interna, basati sui concetti di coesione e separazione dei *cluster*.

L'indice Silhouette Il *Silhouette Coefficient* [69] è un indice che combina le idee della coesione e della separazione e può essere calcolato per ognuna delle osservazioni del dataset, per ciascun *cluster*, oppure per l'intera partizione dei dati. Dato un punto i appartenente al *cluster* C, sia $a(i)$ la distanza media del punto i da tutti i punti appartenenti al cluster C. Si può quindi interpretare $a(i)$ come una misura di quanto il punto i sia dissimile dal suo *cluster* (più piccolo è il valore, migliore è l'assegnazione), nel caso di un intero *cluster* misura di fatto quanto questo sia coeso (più piccolo è il valore, maggiore è la coesione).
Sia $b(i)$ la più piccola distanza media del punto i da ogni altro *cluster* di cui i non è membro. Il *cluster* cui corrisponde il minimo di $b(i)$ è chiamato *neighbouring cluster*, poiché rappresenta dopo C il miglior *cluster* cui assegnare i. Un valore grande di $b(i)$ implica che i sarebbe mal assegnato al suo *neighbouring cluster*. L'indice *Silhouette* per ciascun punto del dataset è ottenuto come segue:

$$s(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]} \tag{3.16}$$

In altre parole al variare di $a(i)$ e $b(i)$ si potrà scrivere:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{se } a(i) < b(i) \\ 0, & \text{se } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{se } a(i) > b(i) \end{cases} \tag{3.17}$$

Dalla precedente espressione è semplice desumere che l'indice *Silhouette*

varia tra -1 e 1. Un $s(i)$ vicino a 1 significa che il dato è ben assegnato al *cluster* C. Con la stessa logica, se il valore di $s(i)$ è vicino a -1, il punto i dovrebbe essere assegnato al suo *cluster* più vicino. Un valore dell'indice *Silhouette* vicino allo zero indica che il dato è al confine di due *cluster* naturali.
Si può validare in maniera individuale ogni *cluster* C_j, con $j = 1, \ldots, k$, usando la media degli $s(i)$ di tutti i punti del *cluster*:

$$s(C_j) = \frac{1}{n_j} \sum_{i \in C_j} s(i) \tag{3.18}$$

L'indice può essere mediato su tutti i punti per calcolare la silhouette dell'intero *Clustering*:

$$S = \frac{1}{k} \sum_{j=1}^{k} C_j \tag{3.19}$$

Il grafico che incrocia i valori assunti dall'indice *Silhouette* al variare di k rappresenta nelle applicazioni uno dei metodi più usati per determinare il numero corretto di *cluster*. Il k ottimo sarà pertanto quello che massimizza il valore di S.

L'indice Dunn L'indice Dunn [23] valuta la qualità di una soluzione di *Clustering* in termini di rapporto tra la separazione e la coesione dei gruppi. Nello specifico questo indice considera la minima distanza a coppie tra i punti appartenenti a differenti *cluster* come misura della separazione tra i *cluster* e il massimo diametro tra tutti i *cluster* come misura della coesione, dove il diametro di un *cluster* è definito come la distanza massima che separa due punti distinti appartenenti allo stesso *cluster* e può essere considerato una misura della dispersione dei diversi *cluster*.

Formalmente:

$$D = \min_{i=1\ldots n_c} \left\{ \min_{j=i+1\ldots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1\ldots n_c}(diam(c_k))} \right) \right\} \tag{3.20}$$

dove: $d(c_i, c_j) = \min_{x \in C_i, y \in C_j} \{d(x, y)\}$ e $diam(c_i) = \max_{x,y \in C_i} \{d(x, y)\}$. Se i *cluster* sono ben separati, ci si aspetta che le distanze tra i gruppi siano elevate e il diametro di ciascun *cluster* sia piccolo [35]. Di conseguenza più alti saranno i valori per il *Dunn Index* migliore risulta lo schema di partizionamento. Le criticità di quest'indicatore sono da ricercarsi nell'elevato costo computazionale e nella sensibilità alla presenza di rumore nei dati (il diametro di un *cluster* può essere sovrastimato in presenza di valori anomali). Le originarie definizioni di *distanza tra cluster* e di *diametro del cluster* sono state generalizzate in [9], dando origine a ben 17 varianti dell'indice Dunn che combinano differenti definizioni delle due misure.

L'indice Calinski-Harabasz L'indice Calinski-Harabasz [13], noto anche come *Variance Ratio Criterion*, valuta la qualità della partizione ottenuta in termini di rapporto tra la varianza entro i gruppi e la varianza tra i gruppi:

$$CH = \frac{\sum_i n_i \times d^2(c_i, c)}{\sum_i \sum_{x \in C_i} d^2(x, c_i)} \times \frac{n-k}{k-1} \tag{3.21}$$

Pertanto potrà anche essere espresso come segue:

$$CH = \frac{trace(\mathbf{B})}{trace(\mathbf{W})} \times \frac{n-k}{k-1} \tag{3.22}$$

dove la traccia della matrice $\mathbf{B}$ è la somma delle varianze tra i gruppi e, analogamente, la traccia della matrice $\mathbf{W}$ è la somma delle varianze entro i gruppi. Anche l'indice CH è un indicatore del tipo $index = a \times separation/b \times cohesion$, dove a e b sono i pesi, pertanto è facile desumere

che più grande sarà il valore del rapporto tra le tracce delle due matrici, migliore sarà la partizione. Il termine $n-k/k-1$ previene che questo rapporto aumenti monotonicamente all'aumentare del numero di *cluster*. Rispetto agli altri indici di validazione interna, quest'indicatore, come anche l'indice Dunn, risulta sensibile alla presenza di rumore nei dati, in quanto in questa circostanza la variabilità entro i gruppi aumenta in maniera più consistente rispetto alla variabilità tra i gruppi. Ciò comporta che l'indice diminuirà a causa dell'influenza del rumore, rendendo così il suo valore instabile [56].

L'indice Davies Bouldin L'indice Davies Bouldin [17] è anch'esso basato sul rapporto tra le distanze entro e tra i gruppi. Nello specifico l'indice valuta la qualità di una data partizione come segue:

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)\right] / d(c_i,c_j)\right\} \quad (3.23)$$

Per ogni *cluster* C, sono calcolate le similarità tra C e tutti gli altri *cluster*, e il valore più alto è assegnato a C come sua similarità. L'indice Davies Bouldin è ottenuto come media delle similarità tra ogni *cluster* e il rispettivo *cluster* più simile ad esso. Si desidera che i *cluster* siano il meno possibile simili l'uno con l'altro, pertanto più piccolo sarà il valore dell'indice migliore sarà la configurazione ottenuta.

L'indice Xie Beni L'indice Xie Beni [89] è principalmente utilizzato per i metodi di raggruppamento di tipo *fuzzy*, ma può essere anche applicato per valutare la qualità di una partizione risultante da metodi di *Crisp Clustering*. Questo indice definisce la separazione tra i *cluster* come la minima distanza al quadrato tra i centri dei *cluster* e la coesione come la media

delle distanze al quadrato tra ogni elemento di ciascun *cluster* e il rispettivo baricentro. Formalmente:

$$XB = \frac{\sum_i \sum_{x \in C_i} d^2(x, c_i)}{n \times min_{i,j \neq i} d^2(c_i, c_j)} \tag{3.24}$$

L'utilizzo di quest'indicatore è ancora una volta finalizzato all'individuazione del numero ottimale di *cluster*, che al variare di k sarà identificato dal minimo valore assunto dall'indice.

L'indice C Per definire l'indice C [38] si considerano le distanze tra le coppie di elementi in ciascun *cluster*. É necessario, quindi, definire le seguenti quantità:

- $N_t = \frac{n(n-1)}{2}$ è il numero totale di coppie di elementi nel dataset;
- $N_w = \sum_{j=1}^{k} \frac{n_j(n_j-1)}{2}$ è il numero totale di coppie di elementi che appartengono al medesimo *cluster*;
- $N_b = N_t - N_w$ è il numero totale di coppie di elementi che appartengono a *cluster* diversi.

L'indice C è definito formalmente attraverso l'equazione 3.25:

$$C = \frac{S_w - S_{min}}{S_{max} - S_{min}} \text{ con } S_{min} \neq S_{max} \tag{3.25}$$

dove: $S_w = \sum_{j=1}^{k} \sum_{i,r \in C_j i<r} d(x_i, x_r)$ è la somma delle N_w distanze tra tutte le coppie di elementi entro il *cluster*. Sostanzialmente per il calcolo dell'indice C si considerano le N_t distanze tra le coppie di elementi come una sequenza di valori ordinati in senso crescente. Le quantità S_{min} e S_{max} sono, quindi, rispettivamente le più piccole e le più grandi N_w distanze tra

tutte le coppie di elementi nell'intero dataset. L'indicatore varia tra 0 e 1, la partizione per cui si ottiene il valore minimo dell'indicatore indica il numero ottimale di *cluster*.

L'indice Gamma L'indice Gamma di Baker-Hubert [3] (equazione 3.26) è una versione riadattata al caso del *Clustering* della statistica Γ di Goodman e Kruskal.

$$G = \frac{s(+) - s(-)}{s(+) + s(-)} \tag{3.26}$$

I confronti sono effettuati tra tutte le dissimilarità entro i *cluster* e tutte quelle tra i *cluster*. Il termine $s(+)$ (rispettivamente $s(-)$) rappresenta il numero di volte in cui la distanza tra due elementi che appartengono allo stesso *cluster* è strettamente più piccola (più grande) della distanza tra due elementi che appartengono a *cluster* diversi. É bene precisare che Il caso in cui ci sia un *ex-aequos* tra i membri dell'equazione non viene considerato nella definizione dell'indicatore [30].

L'indice Point-Biserial In statistica il coefficiente punto-biseriale (Point-Biserial) è una misura della correlazione tra una variabile continua e una variabile binaria. Nell'ambito della *Cluster Validation* l'indice Point-Biserial [61] è una misura della correlazione tra la matrice di dissimilarità e una corrispondente matrice, le cui celle assumono valori pari a 0 o ad 1. Il valore 0 si assegna quando i due corrispondenti elementi appartengono allo stesso *cluster*, si assegna altrimenti un valore pari ad 1. Dato che valori positivi più grandi riflettono un miglior adattamento tra i dati e la partizione ottenuta, il valore massimo dell'indice, al variare del numero di gruppi, corrisponde al numero ottimale di *cluster* nel dataset [62]. Formalmente l'indice è definito

dall'equazione 3.27:

$$PB = \frac{[\bar{S_b} - \bar{S_w}][N_w N_b / {N_t}^2]^{\frac{1}{2}}}{s_d} \tag{3.27}$$

dove:

- $\bar{S_w} = S_w/N_w$ è la media delle N_w distanze tra tutte le coppie di elementi entro il *cluster*;
- $\bar{S_b} = S_b/N_b = \frac{\sum_{j=1}^{k-1}\sum_{l=j+1}^{k}\sum_{i\in C_j r\in C_l} d(x_i,x_r)}{N_b}$ è la media delle N_b distanze tra tutte le coppie di elementi che appartengono a *cluster* diversi;
- s_d è la deviazione standard di tutte le distanze.

L'indice G *plus* L'indice G *plus* [68] è calcolato tramite la seguente equazione:

$$G_p = \frac{2s(-)}{N_t(N_t - 1)} \tag{3.28}$$

Seguendo la notazione utilizzata in precedenza, $s(-)$ rappresenta il numero di volte in cui due elementi che appartengono allo stesso *cluster* presentano una distanza più grande rispetto a quella tra due elementi che appartengono a *cluster* diversi. Al variare del numero di gruppi, il valore minimo dell'indice G *plus* indica il numero ottimale di gruppi.

L'indice Tau L'indice Tau [47] è anch'esso calcolato tramite il confronto tra la matrice di dissimilarità e una seconda matrice con valori 0 e 1 che indica per ciascuna coppia di elementi se gli stessi appartengono o meno al

medesimo *cluster*. Utilizzando la stessa notazione dell'indice Γ l'indice Tau è calcolato nel modo seguente:

$$Tau = \frac{s(+) - s(-)}{N_t(N_t - 1)/2} \tag{3.29}$$

I valori $s(+)$ e $s(-)$ non tengono conto dei legami tra gli elementi, pertanto i confronti in cui la distanza entro il *cluster* è uguale alla distanza tra i *cluster* non rientrano nel numeratore dell'indice. Nell'equazione 3.30 è infatti definìta una versione corretta dell'indice Tau che tiene conto del numero di legami, ottenuta modificando il denominatore:

$$Tau_c = \frac{s(+) - s(-)}{\sqrt{N_b N_w (N_t(N_t - 1)/2)}} \tag{3.30}$$

L'indice McClain Rao L'indice McClain Rao [60] è calcolato come rapporto tra le distanze medie entro e tra i *cluster*:

$$MCR = \frac{\bar{S_w}}{\bar{S_b}} = \frac{S_w/N_w}{S_b/N_b} \tag{3.31}$$

Come già definito, infatti, è la media delle N_w distanze tra tutte le coppie di elementi entro il *cluster* e è la media delle N_b distanze tra tutte le coppie di elementi che appartengono a *cluster* diversi. Il mínimo valore che l'indice assume al variare del numero di gruppi corrisponde al numero ottimale di *cluster* nel dataset.

L'indice Ratkowsky Lance L'indice Ratkowsky Lance [65] è un altro dei criteri proposti in letteratura per la determinazione del numero corretto di gruppi in un dataset. L'indice è formalmente definito nell'equazione 3.32:

$$RL = \sqrt{\frac{\bar{S}}{k}} = \sqrt{\frac{\frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j}}{k}} \tag{3.32}$$

dove:

- $BGSS_j = \sum_k n_k(c_{kj} - \bar{x_j})^2$ è la devianza tra i gruppi di ciascuna variabile x_j con $j = 1 \dots p$;

- $TSS_j = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2$ è la devianza totale di ciascuna variabile x_j.

La quantità al numeratore $\bar{S}$ rappresenta quindi la media dei rapporti tra la devianza *between* e la devianza totale di ogni variabile nella matrice dei dati X. Il numero corretto di gruppi sarà quel valore di k per cui si osserva il valore massimo dell'indicatore.

L'indice Ray Turi L'indice Ray Turi [66] misura la validità complessiva della soluzione di *Clustering* come rapporto tra la compattezza e la separazione dei gruppi ottenuti:

$$RT = \frac{intra}{inter} \tag{3.33}$$

dove:

- $intra = \frac{1}{n}\sum_{j=1}^{k}\sum_{x \in C_j} \|x - c_j\|^2$ è una misura della compattezza dei *cluster* espressa in termini di media delle distanze entro i gruppi;

- $inter = \min_{i<j} \|c_i - c_j\|^2$ è una misura della separazione tra i *cluster* espressa in termini di minima distanza tra i centri dei diversi gruppi.

L'idea di fondo è che minimizzare la distanza entro i *cluster* e al contempo massimizzare la distanza tra i *cluster* equivale a minimizzare il loro rapporto. Pertanto, il valore di k per cui si ottiene il minimo dell'indice corrisponde al numero corretto di gruppi nel dataset.

L'indice Wemmert Gancarski L'indice Wemmert Gancarski [87] misura la qualità del *Clustering* tenendo conto della distanza di ciascun punto della matrice dei dati X dal centro del *cluster* al quale è assegnato e da tutti gli altri $k-1$ centri dei *cluster*. Per ciascun punto x appartenente al *cluster* C_j si calcola infatti il rapporto $R(x)$ (equazione 3.34) tra la distanza del punto considerato dal centro del *cluster* cui appartiene e la più piccola distanza del punto dai centri tutti gli altri *cluster*:

$$R(x) = \frac{\|x - c_j\|}{\min_{l \neq j} \|x - c_l\|} \tag{3.34}$$

Successivamente si considera la media di tali quozienti per ciascun *cluster*. Se la media è maggiore di 1, la stessa viene ignorata, altrimenti si considera il suo complemento ad 1. Questo passaggio è formalmente definito nell'equazione 3.35:

$$J_j = \max \left\{ 0, 1 - \frac{1}{n_j} \sum_{i \in C_j} R(x_i) \right\} \tag{3.35}$$

Infine l'indice Wemmert Gancarski è definito come media ponderata, per tutti i *cluster*, delle quantità J_j, dove i pesi sono rappresentati dal numero n_j di oggetti in ciascun gruppo:

$$WJ = \frac{1}{n} \sum_{j=1}^{k} n_j J_j \tag{3.36}$$

Il valore di k per cui si ottiene il massimo dell'indicatore corrisponde al numero corretto di gruppi nel dataset considerato.

L'indice PBM L'indice PBM è tra gli indici interni di *Cluster Validation* di più recente formulazione [4]:

$$PBM = \left(\frac{1}{k} \times \frac{E_w}{E_t} \times D_b\right)^2 \tag{3.37}$$

dove:

- $E_w = \sum_{j=1}^{k} \sum_{i \in Cj} d(x_i, c_j)$ è la somma delle distanze dei punti di ciascun *cluster* dal rispettivo baricentro;
- $E_t = \sum_{i=1}^{n} d(x_i, c)$ è la somma delle distanze di tutti i punti dal centro c dell'intero dataset;
- $D_b = \max_{i<j} d(c_i, c_j)$ con $i = 1 \dots k$ e $j = 1 \dots k-1$, è una misura della massima separazione tra i *cluster*, espressa in termini di distanza tra i rispettivi centri.

L'indice fornisce una misura della qualità del *Clustering* rispetto a differenti partizioni dei dati. Il massimo valore che assume al variare del numero di gruppi k corrisponde alla migliore partizione dei dati.

L'indice RMSSDT e l'indice RS Gli indicatori presentati finora considerano entrambi i criteri di valutazione (coesione e separazione), in termini di rapporto o di somma, mentre l'indice RMSSDT e l'indice RS [73], insieme all'indice Modified Hubert Γ statistic [39], considerano uno solo degli aspetti. L'indice RMSSDT (Root-mean-square standard deviation) e l'indice RS (R-Squared) sono due indici di validità prevalentemente utilizzati per metodi di *Clustering* gerarchico, ma possono comunque essere usati per valutare i risultati di ogni algoritmo di *Clustering*.
L'indice RMSSDT, formalmente definito nell'equazione 3.38, è la varianza

dei *cluster* quindi misura la loro omogeneità. Essendo l'obiettivo del *Clustering* individuare gruppi omogenei, valori piccoli dell'indice indicano un buon schema di partizionamento.

$$RMSSDT = \sqrt{\frac{\sum_i \sum_{x \in C_i} \|x - c_i\|^2}{P \times \sum_i (n_i - 1)}} \tag{3.38}$$

Trattandosi di un indicatore della variabilità dei *cluster* il minimo valore che potrà assumere sarà zero, tuttavia non abbiamo un *framework* di riferimento per stabilire in ogni altro caso quanto il valore ottenuto sia effettivamente piccolo.
L'indice RS misura la dissimilarità dei *cluster*, nello specifico è calcolato come rapporto tra la somma degli scarti al quadrato tra i *cluster* e la somma degli scarti al quadrato per l'intero dataset, eq. 3.39.

$$RS = \frac{\sum_{x \in D} \|x - c\|^2 - \sum_i \sum_{x \in C_i} \|x - c_i\|^2}{\sum_{x \in D} \|x - c\|^2} \tag{3.39}$$

L'indice RS varia tra 0 e 1, dove 0 significa che non c'è differenza tra i diversi *cluster*, mentre 1 indica che le differenze sono consistenti.

3.3 Altri approcci per la validazione secondo criteri interni

Quando si vuole validare una soluzione di *Clustering*, fornire una valutazione quantitativa e oggettiva della qualità del raggruppamento ottenuto attraverso le misure interne di validazione, è solo uno degli aspetti di cui è possibile tenere conto. Le misure, costruite per valutare la singola soluzione, o per confrontare differenti schemi di *Clustering* senza informazioni *a priori*,

rappresentano infatti degli strumenti utili, ma, come detto, la validazione può essere effettuata perseguendo due ulteriori approcci.
In primo luogo validare una soluzione vuol dire asserire che esiste nei dati una struttura non casuale, in tal senso parleremo di *Cluster Tendency*, nota anche come *Clusterability* [1]. I dati potrebbero non essere "clusterizzabili", violando così l'unica assunzione della *Cluster Analysis*: l'esistenza di una struttura di gruppo. In tal caso, il risultato altro non è che una partizione o una gerarchia, imposta dal metodo di *Clustering* scelto, priva di significato ed inutilizzabile, quale che sia l'obiettivo dell'analisi.
Un ulteriore aspetto da valutare per validare una soluzione di *Clustering* è la stabilità della soluzione stessa. In altre parole si tratta di stabilire quanto è robusta una soluzione sotto perturbazione o sub-campionamento dei dati originali. Una partizione o una gerarchia è considerata stabile quando "cattura" la struttura sottostante un insieme di dati, sotto l'assunzione che tale struttura possa essere riproposta in altri dataset, tratti dalla stessa origine [57]. Facendo uso, ad esempio, di tecniche di ricampionamento, l' obiettivo è comprendere la sensibilità del risultato del *Clustering*, al variare dei parametri di input dell'algoritmo. Pertanto, valutare la stabilità dei risultati di un metodo di *Clustering* rappresenta, oltretutto, uno dei metodi per stimare il numero corretto di *cluster*.

3.3.1 Cluster Tendency

L'obiettivo della *Cluster Tendency* o *Clusterability* è determinare se nella matrice dei dati X ci sono dei gruppi significativi. Il modo più intuitivo per verificare se un insieme di dati presenta dei *cluster* è provare a raggrupparli. Tuttavia, come già sottolineato, quasi tutti gli algoritmi di *Clustering* troveranno comunque dei *cluster* nei dati, indipendentemente dall'effettiva

esistenza di una struttura di gruppo.
Una possibile strategia è analizzare i risultati di un *Clustering* e successivamente affermare che esistono dei gruppi nei dati, nel caso in cui *cluster* ottenuti risultano di buona qualità. Seguendo una strategia di questo tipo, non si tiene conto del fatto che potrebbero esistere dei gruppi di tipo diverso rispetto a quelli individuati dall'algoritmo utilizzato. Per far fronte a questo ulteriore problema si potrebbero confrontare soluzioni di *Clustering* ottenute applicando diversi algoritmi e nel caso in cui i *cluster* risultano uniformemente "poveri", si potrebbe dedurre che non vi siano gruppi nei dati. Queste soluzioni implicano un'analisi a posteriori di una o più soluzioni di *Clustering*.
I metodi di *Cluster Tendency* mirano, invece, a valutare se i dati sono clusterizzabili senza dover applicare alcun algoritmo. L'approccio più comune in quest'ambito è utilizzare dei test statistici di causalità spaziale per valutare quanto i dati si distribuiscono in modo uniforme. Purtroppo, la scelta del modello corretto, la stima dei parametri, e la valutazione della significatività statistica dell'ipotesi che i dati non si distribuiscano casualmente può rivelarsi molto impegnativa.
Un esempio di test di causalità spaziale è la statistica di Hopkins [37]. Si consideri la matrice dei dati X di dimensioni $(n \times p)$ e si generi un campione di m punti, con $m \ll n$. Il campione è generato casualmente dallo stesso spazio dei dati in X. Allo stesso modo, si estragga un campione di ampiezza m direttamente dalla matrice X senza reinserimento. Per ognuno dei due campioni si calcoli la distanza minima di ciascuno dei punti che vi appartengono dai punti nella matrice dei dati d'origine X. Sia u_i la distanza minima di un punto generato dal dataset X, mentre definiamo w_i la distanza minima di ciascun punto estratto dai dati originali in X.

La statistica di Hopkins per i due campioni considerati è definita come segue:

$$H = \frac{\sum_{i=1}^{t} u_i}{\sum_{i=1}^{t} w_i + \sum_{i=1}^{t} u_i} \tag{3.40}$$

Questa statistica paragona la distribuzione dei *nearest-neighbor* dei punti generati in modo casuale, con la stessa distribuzione per il sottoinsieme di punti, campionati casualmente da X. Se il valore di H è prossimo a 0.5 allora le distribuzioni sono simili, il che implica che i dati si distribuiscono casualmente. Valori prossimi a 1 o a 0 indicano rispettivamente che i dati sono altamente clusterizzabili, o che i dati sui distribuiscono regolarmente nello spazio (equi-spaziati). Replicando l'analisi un certo numero di volte e ricalcolando per ciascuna coppia di campioni il valore della statistica di Hopkins H è possibile calcolarne il valore atteso e la varianza, costruendo un intervallo di confidenza per determinare se il dataset X presenta una struttura casuale o meno. Il problema della statistica di Hopkins, e più in generale degli approcci proposti per valutare la *Cluster Tendency* di un dataset, è che risultano applicabili prevalentemente con una ridotta dimensionalità e in spazi euclidei. Inoltre, seppure si riesce ad affermare che i dati sottendono una struttura di gruppo, non abbiamo indicazioni circa la forma o il numero di gruppi presenti.

3.3.2 Cluster Stability

Valutare la stabilità dei risultati di un metodo di *Clustering* rappresenta un ulteriore criterio di validazione interna, in quanto tale valutazione è basata unicamente sui dati a disposizione senza alcun riferimento ad informazioni esterne. L'obiettivo della *Cluster Stability* [57] è asserire quanto sia robusta la soluzione ottenuta sotto perturbazione o sub-campionamento dei dati

originali, per cui, tra gli altri, rappresenta un ulteriore metodo per la determinazione del numero di corretto di *cluster* presenti in un dataset. L'idea di fondo è che le soluzioni ottenute applicando lo stesso schema di *Clustering* su differenti dataset campionati dalla stessa origine (matrice dei dati X) dovrebbero essere simili, riflettendo quindi una certa stabilità. Dato un algoritmo di *Clustering*, un approccio di questo tipo può essere utilizzato per individuare il miglior set di parametri per i dati presi in esame. In particolare, si consideri il caso in cui si vuole individuare il numero corretto di gruppi k nel dataset. É bene precisare che la distribuzione di probabilità congiunta della matrice di dati x è tipicamente ignota. Pertanto, per ottenere un dataset la cui distribuzione sia la stessa dei dati d'origine si possono utilizzare diversi metodi, tra i quali perturbazioni casuali dei dati, sub-campionamento, o ancora tecniche di ricampionamento. Si consideri, ad esempio, una tra le più note tecniche statistiche di ricampionamento è il *Bootstrap* [25], attraverso il quale si generano t campioni di dimensione n estratti con ripetizione dalla matrice dei dati X. Il campionamento con ripetizione implica che uno stesso elemento presente nella matrice X possa essere inserito nel campione più di una volta, pertanto ogni campione generato X_i con $i = 1 \dots t$ sarà differente. Una volta generati mediante il *Bootstrap* t campioni di dimensioni $n \times p$ per ciascuno di questi si esegue lo stesso algoritmo di *Clustering* al variare del numero di gruppi k. Sia $C_k(X_i)$ la partizione ottenuta per un dato valore k del campione X_i e $C_k(X_j)$ la partizione ottenuta per lo stesso valore k del campione X_j. Per valutare la stabilità si confrontano le due partizioni ottenute, utilizzando ad esempio le misure esterne per la validazione (vedi paragrafo 3.2.1) adottate comunemente per il confronto tra una partizione e la classificazione "vera" dei dati, come misure di di distanza o di similarità. Per ogni valore di k a fronte

di tutti i confronti tra le diverse soluzioni è possibile calcolare la distanza a coppie attesa. Il valore di k per il quale si ha la minima deviazione tra le soluzioni ottenute per i dataset ricampionati rappresenta il numero ottimale di gruppi, in quanto presenta la maggiore stabilità. Oltre all'elevato numero di confronti da effettuare un approccio di questo genere presenta comunque delle complicazioni da affrontare. Innanzitutto quando si confrontano le partizioni $C_k(X_i)$ e $C_k(X_j)$ i dataset sottostanti X_i e X_j sono differenti. Prima di confrontare le partizioni è quindi necessario ridurre i vettori delle partizioni $C_k(X_i)$ e $C_k(X_j)$ ai soli elementi in comune a X_i e X_j. Inoltre, poiché i dataset ottenuti attraverso il *Bootstrap* sono generati con un campionamento casuale con ripetizione, uno stesso elemento $x_\alpha \in X$ può presentarsi più volte, pertanto per la costruzione del dataset X_{ij} che contiene gli elementi in comune a X_i e X_j è necessario tener conto anche di questo aspetto. Formalmente il dataset comune X_{ij} è definito come segue:

$$X_{ij} = X_i \cap X_j = \left\{ m^\alpha \,\middle|\, x_\alpha \in X, m^\alpha = \min\left\{ m_i^\alpha, m_j^\alpha \right\} \right\} \quad (3.41)$$

dove m_i^α e m_j^α rappresentano il numero di volte in cui l'elemento x_α si ripete rispettivamente in X_i e X_j.

3.4 L'uso delle misure interne di validazione per la ricerca della partizione ottimale

Kaufmann e Rousseeuw [46] definiscono la *Cluster Analysis* come *"l'arte di trovare gruppi nei dati"*. Da questa definizione si percepisce chiaramente che stimare il numero corretto di gruppi in un dataset è un aspetto cruciale per questo tipo di analisi e non sempre risulta semplice individuare una soluzione. Ad oggi, infatti, non esiste una metodologia condivisa per poter

prendere questa decisione, ed è per questo motivo che l'individuazione della partizione ottimale è stata per anni, e lo è tutt'ora, uno dei problemi di ricerca ancora aperti della *Cluster Analysis* [22][62].
A prescindere dal metodo di *Clustering* scelto, sia esso partitivo o gerarchico, per ottenere una partizione il ricercatore si trova di fronte al problema di dover decidere il numero di gruppi in cui partizionare il dataset preso in esame [26] [75].
Per i metodi di *Clustering* partitivi l'utente deve necessariamente specificare questo parametro prima che venga condotta l'analisi. Al contrario, i metodi gerarchici producono in maniera iterativa una serie di soluzioni, da n classi ad una soluzione in cui tutti gli oggetti appartengono ad un'unica classe (assumendo che gli oggetti nel dataset siano n), senza che debba essere specificato *a priori* il numero di gruppi.
In entrambi i casi, sia se si assegnano valori impropri ai parametri di input dell'algoritmo, sia se la scelta della partizione viene effettuata *ex post* in maniera inadeguata, lo schema di *Clustering* individuato non sarà ottimale per il dataset considerato e ciò comporta, di conseguenza, la possibilità di prendere delle decisioni sbagliate.
In particolare gli errori decisionali che si possono commettere sono sostanzialmente due. Il primo si verifica quando si conclude che nei dati ci sono k gruppi, ma in realtà il numero di gruppi è inferiore a k, pertanto si ottiene una soluzione contenente troppi *cluster*. Il secondo tipo di errore si verifica quando si decide per un numero di gruppi inferiore a quelli effettivamente presenti nei dati, ottenendo così una partizione con pochi *cluster*. Sebbene la gravità dei due tipi di errore cambi a seconda del contesto applicativo, considerare erroneamente un numero troppo esiguo di gruppi comporta una perdita di informazione, derivante dall'unione di gruppi distinti.

Alcuni dei metodi proposti per la ricerca del numero corretto di *cluster* sono piuttosto informali e soggettivi, altri più formali.
Il modo più intuitivo per individuare la partizione ottimale dei dati è la visualizzazione dei risultati. Dopo aver ottenuto una partizione in k gruppi, ad esempio con un metodo partitivo, i dati vengono rappresentati in un diagramma di dispersione, solitamente etichettati con colori diversi secondo i *cluster* di appartenenza. Questo modo di procedere è ampiamente condiviso, anche quando viene presentato un nuovo algoritmo di *Clustering*. Nella maggior parte dei lavori, infatti, si valuta la performance del metodo prendendo come riferimento dataset bidimensionali, allo scopo di rendere il lettore in grado di verificare, attraverso la visualizzazione, la validità dei risultati stessi. Si tratta di una procedura che però non sempre dà risultati illuminanti, soprattutto nel caso di dataset multidimensionali, dove l'effettiva visualizzazione dei dati diventa difficile.
Quando, invece, si conduce una *Cluster analysis gerarchica*, sia utilizzando tecniche agglomerative che divisive, il processo di classificazione produce una gerarchia di partizioni, rappresentata graficamente attraverso il dendrogramma. Questo grafico illustra ad ogni passo dell'algoritmo quali *cluster* sono raggruppati (o divisi) e fornisce il valore della distanza tra questi ultimi prima del raggruppamento. Se la differenza, in termini di distanza, da uno step all'altro dell'algoritmo è elevata, si deduce che i *cluster* che sono stati uniti al successivo step erano relativamente distanti. Questa constatazione implica che il numero corretto di *cluster* sarà quello che precede un grande salto in termini di distanza [28]. Il vantaggio dei metodi gerarchici è proprio la possibilità di visualizzare in un unico grafico le diverse partizioni contemporaneamente e scegliere *ex post* il numero di gruppi. Tuttavia nel caso di dati altamente dimensionali si preferisce in genere un approccio partitivo,

in quanto i metodi gerarchici sono caratterizzati da un'elevata complessità computazionale.

Un approccio più formale per la determinazione del numero corretto di gruppi è quello di valutare statisticamente la bontà della soluzione di *Clustering* ottenuta. In tal caso è possibile applicare test statistici per verificare se essa è significativamente diversa da una ottenibile per caso. In genere si verifica se la distanza tra le medie dei gruppi è significativa. Per stabilire la significatività dell'applicazione di tecniche che producono partizioni ottimizzando le funzioni della matrice di devianze-codevianze, è possibile utilizzare la statistica lambda di Wilks [27]. L'uso di test per la verifica di ipotesi in questo contesto è piuttosto dibattuto. Infatti, al crescere dell'ampiezza n del collettivo, i centroidi dei gruppi risultanti da qualsiasi procedimento classificatorio risultano quasi sempre significativamente diversi.

L'uso delle misure interne per la validazione dei *cluster* consente di identificare la soluzione che produce una maggiore discontinuità tra i gruppi identificati e una maggiore omogeneità all'interno dei gruppi. Nello specifico si osserva il grafico che incrocia i valori assunti dalla misura di validazione scelta al variare del numero di *cluster*. Di solito si esegue un algoritmo di *Clustering* un numero di volte pari all'intervallo di valori considerato per il numero di gruppi $k_{min} < k < k_{max}$, e per ciascuna di queste soluzioni viene calcolato l'indice di validazione. Si sceglierà il valore di k per cui si ottiene il valore migliore per il criterio di validazione scelto.

A seconda, quindi, dell'indice di validazione utilizzato, la partizione migliore dei dati sarà quella corrispondente al massimo o al minimo valore dell'indicatore, o ancora quella per cui si osserva nel grafico un punto di flesso, detto anche *knee* o *elbow*. Il punto di flesso nel grafico corrisponde alla più grande differenza tra due salti successivi e consente di individuare il miglior

valore per k.
Più precisamente definiamo $V_i = Q_{i+1} - Q_i$ la pendenza tra due punti successivi nel diagramma, dove Q_{i+1} e Q_i sono i valori assunti dalla misura di validazione in corrispondenza rispettivamente delle partizioni in $i+1$ e in i gruppi. Il numero corretto di gruppi k, è:

$$k = \arg \max_{k_{min}<k<k_{max}} (V_i - V_{i-1}) \tag{3.42}$$

Questo concetto può essere più semplicemente spiegato osservando il grafico in Fig. 3.2. Come si può notare la rappresentazione grafica dei dati suggerisce la presenza di 4 *cluster*. L'indice di validazione viene calcolato per le partizioni da 2 a 7 *cluster*.

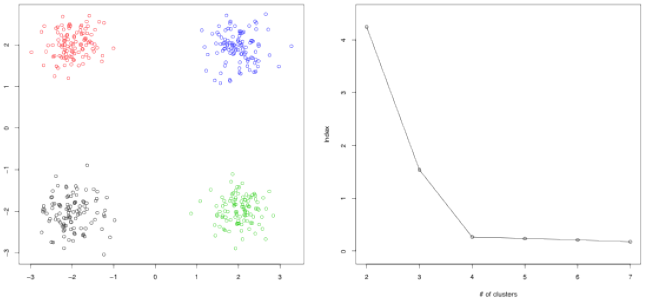


Figura 3.2: Esempio di *elbow*

Osservando il grafico a destra è possibile osservare l'*elbow* in corrispondenza della partizione in 4 gruppi, confermando l'appartenenza dei dati ad esattamente 4 gruppi distinti. Tuttavia non sempre questo tipo di approccio funziona bene. I *cluster* possono essere considerevolmente sovrapposti, a differenza di quelli in Fig. 3.2, i dati possono contenere *cluster* annidati.
Inoltre, spesso se si utilizza più di una misura di validazione si ottengono

risultati discordanti.

Allo scopo di comprendere il comportamento dei diversi indici interni di validazione nel prossimo capitolo sono presentati i risultati di un'analisi comparativa condotta su un insieme di 15 indicatori differenti.

Capitolo 4

Strategie di validazione per il Clustering di documenti

In un processo di *Clustering*, in cui non è nota *a priori* alcuna informazione circa la composizione dei gruppi, i criteri interni per la validazione rappresentano l'unica opzione disponibile per valutare la qualità della soluzione. La validazione della soluzione di *Clustering* può avvenire seguendo diverse strade: cercando di valutare a monte se i dati siano effettivamente raggruppabili (*Cluster tendency*); valutando la stabilità dell'algoritmo rispetto a differenti campioni della base di dati d'origine, con il vantaggio di essere indipendenti dal metodo di *Clustering* scelto, pur non consentendo però una validazione diretta della soluzione ottenuta (*Cluster stability*); o ancora, facendo riferimento a misure che valutino la compattezza e la separazione dei gruppi individuati (*Cluster validation*). Date le differenze tra i tre approcci menzionati non risulta possibile compararli riferendosi ad uno stesso quadro di riferimento, pertanto l'attenzione è stata focalizzata unicamente sulle misure interne di *Cluster validation*. Sebbene gli altri due approcci siano

ugualmente validi comportano una serie di operazioni non banali per ottenere un risultato definitivo circa la qualità della soluzione ottenuta, mentre le misure interne di *Cluster validation* sono strumenti in grado di fornire, attraverso una valutazione diretta un'informazione sintetica circa tale qualità, in particolare in termini di numero di gruppi presenti nel dataset.
In questo capitolo sono, quindi, presentati i risultati di uno studio comparativo sugli indici interni di validazione in modo da descriverne criticamente il funzionamento e le specificità. Sono proposti i risultati ottenuti con differenti algoritmi di *Clustering* sia su dati simulati sia su dati reali di tipo quantitativo. A partire da tale studio sono poi state analizzate le performance degli indici con gli algoritmi più idonei per il *Clustering* di documenti, in modo da poter approntare un secondo studio comparativo su basi di dati di tipo testuale.

4.1 Analisi comparativa per la scelta di una misura di validazione

Nell'idea di fornire una misura di affidabilità delle soluzioni ottenute applicando metodi di *Clustering* a problemi reali, tra i diversi approcci disponibili per la validazione secondo criteri interni si è scelto di prediligere gli indici interni di *Cluster Validation*. Come detto, la scelta di misure interne di validazione è dettata dalla necessità di misurare sinteticamente ed in maniera oggettiva la qualità della soluzione ottenuta, con particolare attenzione alla determinazione del numero corretto di *cluster* presenti in un dataset. É bene precisare che le misure interne di validazione sono sostanzialmente degli indicatori assoluti, pertanto non si ha alcuna informazione per ciò che concerne il rispettivo campo di variazione. Se ad esempio otteniamo per

un indicatore un valore pari a 5, questo rappresenta un risultato buono, discreto o mediocre?
L'individuazione di un *framework* di riferimento per interpretare i valori ottenuti è, quindi, un aspetto cruciale, in quanto questo problema implica la difficoltà di eseguire un confronto tra differenti indici.
Seguendo un approccio ampiamente condiviso in letteratura si calcola il valore assunto dall'indice al variare dei parametri di input dell'algoritmo considerato (numero di *cluster*, numero di *nearest neighbours*, etc...), individuando così la soluzione di *Clustering* "migliore" in quella che ottimizza il criterio sulla base del quale l'indice è stato costruito.
Questo stesso modo di procedere è adottato per confrontare i risultati ottenuti dall'applicazione sullo stesso set di dati di differenti algoritmi, che il più delle volte forniscono però soluzioni discordanti. La differenza tra le soluzioni non è una sorpresa: i diversi algoritmi cercano infatti soluzioni diverse, poiché presuppongono a monte una diversa definizione di "*cluster*", che di conseguenza si traduce in un diverso criterio statistico da ottimizzare. Risulta quindi impossibile individuare un unico indice che sia in grado di valutare la qualità di una soluzione indipendentemente dall'algoritmo che l'ha generata. Inoltre molti degli algoritmi di *Clustering* sono stati proposti proprio come soluzioni per raggruppare dati caratterizzati da una particolare struttura. In altre parole, a seconda della struttura dei dati un algoritmo sarà evidentemente più efficace di altri nell'individuare i *cluster* (naturali).
L'obiettivo dell'esperimento di seguito presentato è effettuare uno studio comparativo che consenta di analizzare il comportamento degli indicatori in un gran numero di configurazioni, allo scopo di individuare quali di essi forniscano una risposta più adeguata rispetto alla naturale partizione dei dati.

4.1.1 Setup dell'esperimento

L'esperimento è stato pianificato allo scopo di confrontare i diversi indicatori in un'ampia varietà di configurazioni, tenendo conto di diversi fattori. Sfortunatamente a causa dell'esplosione delle combinazioni possibili, ciascuno di questi fattori è necessariamente limitato a pochi livelli e produce un confronto tra numerose configurazioni alternative.
La metodologia comparativa adottata è ampiamente condivisa in letteratura per la valutazione della qualità di una soluzione di *Clustering* ed in particolare per la determinazione del corretto numero di gruppi in un dataset. Come detto, la strategia consiste nell'eseguire un algoritmo per un insieme di k differenti valori del numero di gruppi su uno specifico dataset, ottenendo così un insieme di partizioni. Si calcola poi il valore assunto dalla misura interna di validazione per tutte le partizioni ottenute.
Il numero di *cluster* nella partizione, alla quale corrispondono i migliori risultati, è considerato la predizione dell'indice di validazione. Nello specifico, tale predizione sarà soddisfacente se il numero di gruppi individuato dall'indicatore coincide con il vero numero di classi del dataset considerato.
La strategia è stata condotta su 12 dataset simulati disponibili sul sito *Speech and Image Processing Unit* (http://cs.joensuu.fi/sipu/data sets/) dell'*University of Eastern Finland*, ma comunque facilmente reperibili in rete. Nella Tabella 4.1 sono riportate le caratteristiche di ciascuno di questi dataset, tenendo conto del numero di oggetti, del numero di dimensioni e del numero di classi note *a priori*.
Tutti i dataset sono simulati a partire da una distribuzione Gaussiana.
Inoltre, essendo il fine ultimo dell'esperimento incentrato in particolar modo sull'individuazione di misure che siano in grado di fornire un risultato

Tabella 4.1: Descrizione dei dataset simulati

Dataset	**N.Oggetti**	**Dimensioni**	**N.Classi**
Aggregation	788	2	7
Dim032	1024	32	16
Dim064	1024	64	16
Dim128	1024	128	16
Dim256	1024	256	16
Dim512	1024	512	16
Dim1024	1024	1024	16
R15	600	2	15
S2	5000	2	15
Flame	240	2	2
Pathbased	300	2	3
Jain-Toy	373	2	2

attendibile circa la naturale partizione dei dati nelle applicazioni reali, sono stati presi in considerazione 9 dataset reali. Si tratta di dataset (Tabella 4.2) noti in letteratura, scaricati dal sito *UC Irvine Machine Learning Repository* (https://archive.ics.uci.edu/ml/index.html).

É importante sottolineare che tutti i dataset qui presentati sono solitamente utilizzati nell'ambito di problemi di Apprendimento supervisionato, come ad esempio la valutazione di algoritmi di Classificazione, e quindi non sempre adattabili al problema del *Clustering*.
Su ciascuno dei dataset presi in considerazione sono stati utilizzati 3 diversi algoritmi di *Clustering* per il calcolo delle partizioni, di cui uno è appartenente alla famiglia degli algoritmi partitivi e due a quella degli algoritmi gerarchici.
Per quanto riguarda la prima famiglia di metodi è stato scelto il *Kmeans*

Tabella 4.2: Descrizione dei dataset reali

Dataset	**N.Oggetti**	**Dimensioni**	**N.Classi**
Glass	214	9	6
Iris	150	4	3
Wine	178	13	3
Yeast	1484	8	10
Breast tissue	106	9	6
Ecoli	336	7	8
Winequalityred	1599	11	6
Breast Wisconsin	569	30	2
Transfusion	748	4	2

comunemente preso a riferimento negli studi di letteratura, rappresentando ad oggi l'algoritmo più utilizzato nelle applicazioni di *Clustering*.
Riferitamente ai metodi gerarchici, sono stati scelti due algoritmi agglomerativi: l'*Average linkage Clustering* e l'algoritmo di *Ward* [85]. La differenza sostanziale tra i due metodi è il modo in cui i diversi *cluster* sono aggregati ad ogni passo dell'algoritmo.
Nell'*Average linkage Clustering* la distanza tra due diversi *cluster*, come visto nel capitolo precedente, è definita in termini di distanza media tra le coppie di oggetti appartenenti ai due diversi gruppi. L'algoritmo di *Ward* opera invece in maniera diversa ai fini delle aggregazioni successive, in quanto si basa sulla scomposizione della devianza totale nella somma delle devianze entro i *cluster* e tra i *cluster*. Oltre ad essere degli algoritmi ben noti e tra i più tradizionali nell'ambito del *Clustering*, è possibile ottenere facilmente da ciascuno di essi differenti partizioni modificando il parametro che controlla il numero di *cluster* della partizione finale.
Per ciascuno dei precedenti dataset ogni algoritmo è stato eseguito un nu-

mero di volte compreso tra 2 e $2g$, dove g è il numero "vero" di gruppi nello specifico dataset. Per i dataset reali per i quali il numero di gruppi "vero" è pari a 2 abbiamo scelto come estremo superiore dell'intervallo 6, calcolando così almeno 5 partizioni per ogni algoritmo.
Per poter confrontare le performance delle diverse misure interne di validazione sono stati selezionati complessivamente 15 indici. Nella tabella 4.3 sono riportati gli indici utilizzati, classificati secondo il criterio di ottimizzazione (min o max) e la logica di costruzione (*graph-based* o *prototype-based*). Considerando sia i dataset simulati che quelli reali e tenendo conto

Tabella 4.3: Indici interni di validazione

	Prototype-based	**Graph-based**
Min	Xie-Beni (XB) Ray Turi (RT) Davies-Bouldin (DB)	McClainRao (MCR) C Index (C) G+ (G_p)
Max	Calinski Harabasz (CH) Ratkowski Lance (RL) PBM Wemmert Gancarski (WG)	Gamma (G) Tau (T) Point Biserial (PB) Dunn (D) Silhouette (S)

del differente numero di partizioni ottenute per ognuno di essi, i 15 indici di validazione sono stati calcolati complessivamente per tutte le partizioni.

4.1.2 Risultati

Nelle tabelle qui di seguito illustrate sono riportati in sintesi i risultati ottenuti dalle diverse analisi comparative, condotte sia sui dataset simulati sia su quelli reali.
Ciascuno dei valori degli indici di validazione calcolati per le diverse partizio-

ni è espresso in termini di rango. Un valore pari a 1 indica in particolare che l'indice ha individuato come miglior soluzione possibile tra tutte le $(2g-1)$ partizioni quella vera, mentre un valore più elevato (in ordine crescente) implica un maggior allontanamento dalla vera classificazione dei dati. Si rimanda comunque all'Appendice A per i risultati completi.
Osservando innanzitutto i dataset simulati è possibile evidenziare come nel caso del *Kmeans* gli indici di validazione, sia quelli *graph-based* sia quelli *prototype-based*, abbiano fornito in generale dei risultati scoraggianti.

Tabella 4.4: Kmeans dataset simulati

		GRAPH-BASED								**PROTOTYPE-BASED**						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Aggregation	7	1	2	1	8	7	5	3	6	3	5	4	6	2	4	5
Dim032	16	16	16	18	18	19	19	15	9	22	18	30	18	27	22	24
Dim064	16	5	13	14	13	14	11	15	9	10	23	8	10	15	3	16
Dim128	16	8	12	12	15	12	12	15	11	13	19	15	8	15	14	16
Dim256	16	4	12	11	9	9	9	15	12	8	4	8	6	11	1	15
Dim512	16	9	13	18	16	16	16	15	7	14	10	13	10	20	10	22
Dim1024	16	5	11	10	10	9	9	15	15	8	10	10	9	11	8	13
R15	15	15	20	13	16	14	11	9	11	14	4	7	14	24	9	17
S2	15	9	26	12	15	16	14	2	14	5	3	2	14	18	2	25
Flame	2	5	5	5	5	5	5	5	5	5	5	4	2	5	5	5
Pathbased	3	1	3	1	5	4	1	1	1	2	1	4	1	1	1	3
Jain-Toy	2	5	2	5	5	5	2	1	3	5	4	5	1	4	1	5

Diversamente, nel caso dell'*Average linkage Clustering* (UPGMA) e dell'algoritmo di *Ward* i risultati sono in taluni casi molto buoni, in particolar modo nel caso degli indici *prototype-based*. La diversa dimensionalità dei dataset non sembra influire sulle performance degli algoritmi e quindi degli indici. Sebbene sia ampiamente diffusa e condivisa l'idea che il *Kmeans* dia risultati significativi quando i cluster hanno livelli di densità e numerosità non troppo dissimile, e soprattutto forma sferica, l'esperimento condotto sembra fornire risposte contraddittorie. Indubbiamente le scelte effettuate in merito alla procedura utilizzata per la simulazione influenzano i risultati.

Tabella 4.5: Upgma dataset simulati

		GRAPH-BASED								**PROTOTYPE-BASED**						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Aggregation	7	2	9	2	8	8	5	1	5	9	8	3	6	5	3	10
Dim032	16	1	1	1	17	16	1	1	14	1	15	1	15	1	1	1
Dim064	16	1	1	1	17	16	1	1	14	1	3	1	14	1	1	1
Dim128	16	1	1	1	17	11	1	1	14	1	13	1	11	1	1	1
Dim256	16	1	1	1	17	4	1	1	14	1	10	1	9	1	1	1
Dim512	16	1	1	1	17	13	1	1	14	1	1	1	9	1	1	1
Dim1024	16	1	1	1	17	1	1	1	14	1	1	1	1	1	1	1
R15	15	1	13	4	16	14	13	5	13	1	5	1	14	1	1	13
S2	15	11	22	15	15	15	14	3	14	1	3	1	14	2	1	19
Flame	2	5	5	5	4	5	5	5	5	5	5	5	2	4	5	5
Pathbased	3	4	4	4	5	4	1	3	2	1	4	5	1	1	1	4
Jain-Toy	2	5	3	5	3	5	5	3	5	5	5	5	1	3	5	5

Tabella 4.6: Ward dataset simulati

		GRAPH-BASED								**PROTOTYPE-BASED**						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Aggregation	7	8	6	7	8	8	5	7	6	4	8	5	6	9	5	10
Dim032	16	1	1	1	17	16	1	1	14	1	1	1	15	1	1	1
Dim064	16	1	1	1	17	16	1	1	14	1	1	1	14	1	1	1
Dim128	16	1	1	1	17	11	1	1	14	1	10	1	12	1	1	1
Dim256	16	1	1	1	17	7	1	1	13	1	1	1	11	1	1	1
Dim512	16	1	1	1	17	13	1	1	14	1	1	1	8	1	1	1
Dim1024	16	1	1	1	17	1	1	1	13	1	1	1	1	1	1	1
R15	15	1	8	1	16	16	13	5	13	1	8	1	14	1	1	7
S2	15	1	1	1	16	16	14	2	14	1	1	1	14	1	1	1
Flame	2	5	5	5	4	5	5	5	5	5	5	5	3	5	5	5
Pathbased	3	1	2	1	5	4	1	2	1	1	1	4	1	1	1	3
Jain-Toy	2	5	2	5	5	5	1	1	3	5	5	5	1	2	1	5

Tabella 4.7: Kmeans dataset reali

		GRAPH-BASED								**PROTOTYPE-BASED**						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Glass	6	2	4	2	11	2	4	6	2	2	7	1	4	4	4	4
Iris	3	4	2	4	4	4	2	2	2	1	4	1	2	2	2	3
Wine	3	1	3	1	5	1	1	1	1	1	1	1	1	2	3	3
Yeast	10	11	14	11	11	11	8	8	9	8	11	6	8	3	8	13
Breast tissue	6	5	6	5	9	7	5	5	4	3	3	3	4	6	5	6
Ecoli	8	9	14	9	9	9	7	11	7	7	9	6	7	15	13	15
Winequalityred	6	4	6	1	7	7	3	3	2	3	3	1	5	3	3	8
Breast Wisconsin	2	4	1	1	4	5	1	1	1	4	2	5	1	1	1	1
Transfusion	2	5	1	3	3	4	1	3	4	4	4	4	1	1	1	2

Tabella 4.8: Upgma dataset reali

		GRAPH-BASED								PROTOTYPE-BASED						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Glass	6	11	2	12	5	12	8	5	8	11	10	8	9	13	13	9
Iris	3	5	5	5	4	4	2	2	2	1	4	1	2	3	2	5
Wine	3	4	5	4	2	4	4	2	4	4	2	1	4	4	4	5
Yeast	10	6	3	7	7	7	10	6	11	17	8	8	10	8	9	4
Breast tissue	6	8	9	8	9	8	5	5	1	6	7	6	5	6	5	10
Ecoli	8	7	4	7	9	7	1	7	1	2	9	6	3	4	2	13
Winequalityred	6	5	7	8	5	8	3	5	5	7	1	5	4	8	5	9
Breast Wisconsin	2	4	2	2	1	4	5	1	4	4	5	5	3	1	1	2
Transfusion	2	2	3	2	1	1	5	2	5	1	5	1	1	1	1	2

Tabella 4.9: Ward dataset reali

		GRAPH-BASED								PROTOTYPE-BASED						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
Glass	6	4	4	5	11	2	5	7	5	2	9	1	4	6	7	6
Iris	3	5	5	3	4	4	2	2	2	1	4	1	2	2	2	5
Wine	3	2	1	3	4	4	1	2	1	1	1	2	1	2	1	2
Yeast	10	11	11	11	11	11	5	9	7	9	10	8	6	17	8	11
Breast tissue	6	6	3	6	11	8	4		1	7	6	6	4	3	4	3
Ecoli	8	10	10	9	9	9	5	13	6	7	15	7	7	13	7	12
Winequalityred	6	7	4	5	8	7	2	1	2	5	7	5	2	1	1	2
Breast Wisconsin	2	3	1	1	1	4	1	1	3	5	1	5	1	1	1	1
Transfusion	2	5	5	5	4	5	3	5	4	4	5	5	1	3	1	5

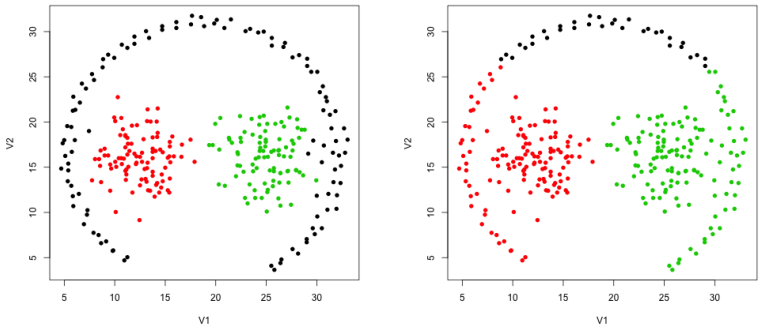
Nel caso dei dataset reali la situazione sembra essere diametralmente capovolta rispetto a quanto detto in precedenza. Si può infatti osservare come nel caso del *Kmeans* i risultati siano migliori rispetto a quanto ottenuto con gli altri due algoritmi presi in considerazione. In particolare l'utilizzo di indici *prototype-based* nel primo caso sembra essere la miglior scelta. Questo conferma l'idea che nella scelta degli indici di validazione è bene tener sempre presente la natura dell'algoritmo di *Clustering* utilizzato, anche se in generale gli indici *prototype-based* sembrano essere una scelta convincente sia per algoritmi partitivi che gerarchici.
Tuttavia ad un più attento esame è possibile notare come la stessa procedura di confronto degli indici adottata comunemente nella letteratura della *Cluster Validation* evidenzi delle criticità. A scopo di esempio in Tabella 4.10 si riportano innanzitutto i risultati del *Kmeans* relativi al dataset simulato *Pathbased*. Come è possibile osservare guardando ai ranghi la maggior parte degli indici suggerisce come migliore partizione quella in 3 gruppi. Questo risultato è apparentemente in accordo con il numero vero di gruppi nel dataset considerato. Se però ci si sofferma sulla partizione in 3 gruppi

Tabella 4.10: Dataset Pathbased - Kmeans

Pathbased	GRAPH-BASED								PROTOTYPE-BASED						
	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
kmeans.2	5	1	5	4	5	2	2	2	1	2	1	2	2	2	1
kmeans.3	1	3	1	5	4	1	1	1	2	1	4	1	1	1	3
kmeans.4	4	4	4	3	3	3	3	3	3	5	2	3	4	3	4
kmeans.5	3	2	3	2	2	4	4	4	5	4	3	4	5	4	2
kmeans.6	2	5	2	1	1	5	5	5	4	3	5	5	3	5	5

ottenuta con il *Kmeans* e la si confronta con le vere etichette del dataset si evince chiaramente che, seppure gli indici forniscono una risposta adeguata in termini di numero di gruppi, la partizione ottenuta non riflette la naturale

struttura dei dati. Come evidenziato in Fig. 4.1, la composizione dei gruppi è infatti completamente differente nei due casi. É bene inoltre precisare che una situazione analoga si verifica anche per gli altri due metodi di *Clustering* considerati. In Tabella 4.11 sono presentati i risultati del *Kmeans* relativi al



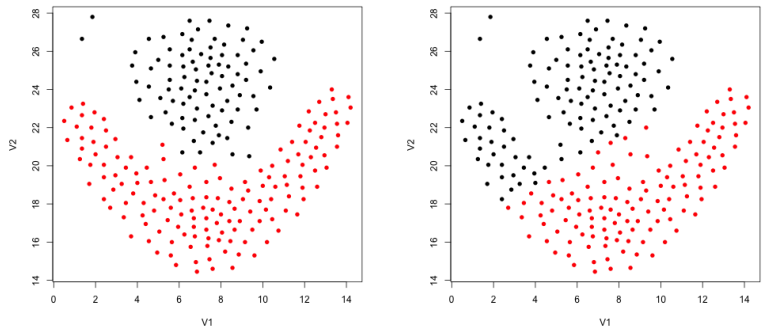
(a) classificazione vera (b) Kmeans k=3

Figura 4.1: Vera classificazione e partizione per il dataset Pathbased

dataset simulato *Flame*. Nessuno degli indicatori individua come migliore partizione quella in 2 gruppi, mentre in realtà il numero vero di gruppi nel dataset *Flame* è pari proprio a 2.

Tabella 4.11: Dataset Flame - Kmeans

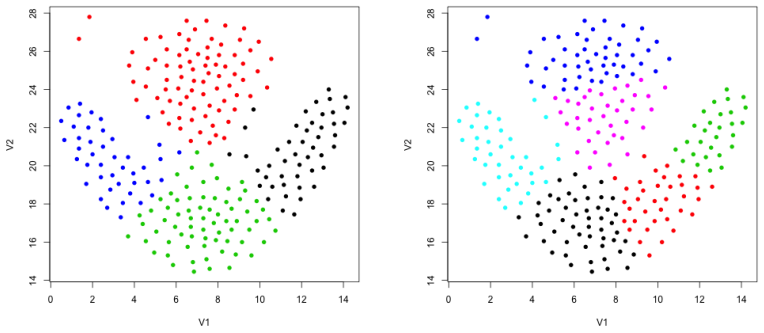
Flame	GRAPH-BASED								PROTOTYPE-BASED						
	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
kmeans.2	5	5	5	5	5	5	5	5	5	5	4	2	5	5	5
kmeans.3	4	4	4	4	4	2	2	2	4	3	5	1	2	2	4
kmeans.4	2	2	2	3	3	1	1	1	1	1	1	3	1	1	3
kmeans.5	3	3	3	2	2	3	3	3	3	2	2	4	4	3	2
kmeans.6	1	1	1	1	1	4	4	4	2	4	3	5	3	4	1



(a) classificazione vera (b) Kmeans k=2

Figura 4.2: Vera classificazione e partizione per il dataset Flame

In ogni caso confrontando le vere etichette del dataset con la partizione del *Kmeans* in 2 gruppi (Fig. 4.2) si evince come anche in questo caso l'algoritmo non è in grado di individuare la struttura sottostante i dati. Questa non è certo una sorpresa in quanto i dati stessi sono caratterizzati da una struttura difficilmente rilevabile in maniera automatica. É interessante osservare come, invece di una partizione in 2 gruppi, circa una metà degli indici suggeriscono una partizione in 4 gruppi (Fig.4.3 (a)), mentre altrettanti indicatori individuano come migliore partizione quella in 6 gruppi (Fig.4.3 (b)). In realtà gli indicatori, in coerenza con la logica in base alla quale sono costruiti, misurano di volta in volta la coesione e la separazione dei gruppi ottenuti, pertanto così come si evince dai risultati presentati rispetto ad una partizione con un numero più contenuto di gruppi prediligono in tal caso una partizione in più gruppi che siano al contempo più coesi e meglio separati.



(a) Kmeans k=4 (b) Kmeans k=6

Figura 4.3: Flame - Classificazione suggerita dagli indici

4.2 Analisi comparativa su dataset testuali

Come già detto in precedenza, in assenza di informazione *a priori* è quasi scelta obbligata affidarsi a tecniche di *Clustering* per la categorizzazione dei documenti. Il dibattito sulla scelta del miglior criterio da adottare, e conseguentemente dell'algoritmo da scegliere, è tuttora vivo e acceso nella letteratura. Moltissime sono le soluzioni che provengono dall'ambito Informatico, ma spesso sembra che l'obiettivo principale sia quello di ottenere soluzioni in tempi di calcolo sempre più ristretti, tenuto conto della complessità computazionale richiesta, piuttosto che tener conto di accuratezza e significatività statistica. Sono ben pochi i lavori che si preoccupano ad esempio di effettuare, al termine dei confronti su dataset noti e diffusi nella comunità scientifica, dei test per verificare le effettive migliori *performance* degli algoritmi via via proposti rispetto a quelli esistenti. Inoltre potenziali distorsioni ben note nel caso dei dati tradizionali sembrano essere anche

più ponderose nel trattamento di testi scritti in linguaggio naturale. Gli algoritmi gerarchici spesso inseriscono documenti dello stesso tipo – in termini di similarità – nello stesso gruppo fin dai primi passi del processo di *Clustering*. Ciò implica che documenti simili per vocabolario utilizzato ma differenti per categoria di appartenenza sono frequentemente accoppiati. A causa della logica di funzionamento degli algoritmi gerarchici agglomerativi tali "errori" non possono essere corretti negli step successivi, limitando quindi l'accuratezza dei risultati. Per contro il processo è completamente automatico, e al ricercatore è lasciato solamente il compito di decidere a quale livello tagliare l'albero gerarchico e quindi quale partizione conservare. Quest'ultimo passaggio è comunque confortato dall'utilizzo di strumenti quali quelli presentati nel paragrafo 3.3. Per quanto detto i metodi partitivi sono sicuramente preferibili rispetto a quelli gerarchici nell'ambito del *Text Clustering*, sebbene non sia necessariamente altrettanto vero in altri domini particolari. Pur avendo la necessità di una inizializzazione, e quindi della scelta di determinati parametri di funzionamento, gestiscono meglio la complessità computazionale dovuta alle notevoli dimensioni degli oggetti in termini di caratteristiche (*feature*) utilizzate per descriverli. Nel caso in cui non si può ritenere attendibile la similarità tra documenti, è necessario avere a riferimento proprietà più globali. Calcolare la similarità di un documento da un centroide equivale più o meno a calcolare una similarità media rispetto a tutti i documenti già assegnati ad un *cluster*. Sono stati comunque proposti diversi approcci per combinare le due diverse prospettive e prendere quindi le caratteristiche migliori di entrambi. Un possibile modo di procedere ad esempio potrebbe essere quello di utilizzare prima un metodo gerarchico in modo da definire il numero delle classi, risolvendo così il problema della inizializzazione degli algoritmi partitivi, e quindi di

migliorare i risultati ottenuti con l'utilizzo di un algoritmo partititivo. Alla luce delle considerazioni fin qui sviluppate di seguito l'esperienza fatta nel paragrafo 4.1 è ripetuta con riferimento ai dati testuali.

4.2.1 Setup dell'esperimento

La strategia seguita per l'analisi comparativa degli indici interni di validazione, condotta su dataset quantitativi sia simulati che reali, è parimenti riportata al caso in cui i dati da raggruppare sono documenti.
Sono stati selezionati 13 *corpora* testuali tra quelli maggiormente utilizzati in letteratura per il confronto di algoritmi di *Text Clustering* e di *Text Classification*.
Il *corpus fbis* è stato ottenuto a partire dai documenti del *Foreign Broadcast Information Service* contenuti nella collezione TREC-5, e le classi corrispondono alle categorie utilizzate nella collezione. I *corpora tr11*, *tr12*, *tr23*, *tr41* e *tr45*, sono stati derivati dalle collezioni TREC-5, TREC-6 a TREC-7, e le classi utilizzate corrispondono ai documenti che sono stati giudicati rilevanti in base a particolari query.
I *corpora la1* e *la2* sono stati ottenuti dagli articoli del Los Angeles Times usati in TREC-5; le categorie utilizzate includono documenti appartenenti alle rubriche di Finanza, Affari Esteri, Cronaca Nazionale, Cronaca Locale, Entertainment e Sport.
I *corpora k1a*, *k1b* e *wap* sono parte del progetto *WebACE*; i documenti corrispondono alle pagine web elencate nelle directory per argomento di *Yahoo!*, categorizzati secondo un diverso livello di granularità.
Su tutti i *corpora* è stato effettuato il medesimo pretrattamento, allo scopo di trasformare l'informazione non strutturata contenuta nelle diverse raccolte di documenti in dati strutturati analizzabili con tecniche proprie di

analisi statistica. Per il pretrattamento dei documenti è stato effettuato in prima istanza il *parsing* per l'individuazione delle diverse forme grafiche e conseguentemente la costruzione del vocabolario di riferimento.. Le successive operazioni di pulizia del testo sono state la rimozione dei numeri e della punteggiatura, e di eventuali spazi aggiuntivi tra le diverse forme grafiche. Sono state poi normalizzate le forme grafiche provvedendo all'abbassamento delle maiuscole in maniera tale da ridurre la dimensione del vocabolario della raccolta attraverso il giusto *matching* tra forme grafiche identiche ma sostanzialmente scritte in modo diverso, sebbene questa operazione possa comportare al contempo un'aumento dell'ambiguità dei termini nella raccolta.
Infine prima di procedere alla costruzione delle tabelle lessicali, i documenti di ciascuna raccolta sono stati lemmatizzati attraverso l'utilizzo di un lemmatizzatore probabilistico, allo scopo di ricondurre le diverse forme grafiche già normalizzate, al rispettivo lemma di appartenenza. Inoltre, dal vocabolario di lemmi così ottenuto sono successivamente state eliminate attraverso una *stop-list* standard della lingua inglese le parti del discorso funzionali, quali articoli, preposizioni, avverbi.
É bene precisare che si è preferito seguire la pratica comune delle applicazioni di *Text Mining*, evitando di effettuare operazioni brutali di pretrattamento. Ragionando in termini di similarità non si è ritenuto opportuno prediligere un'unica categoria grammaticale, considerando ad esempio solo i sostantivi, o ancora eliminando termini che si presentano sopra una certa soglia di frequenza, in quanto si potrebbero forzatamente eliminare, termini che fanno parte di un vocabolario comune a diversi documenti. La codifica dell'informazione testuale contenuta in ciascuna raccolta è stata effettuata seguendo lo schema BOW scegliendo come sistema di pesi il *Term Frequen-*

cy, pertanto ognuna delle tabelle lessicali di dimensione $(n \times p)$, dove n è il numero di documenti nella raccolta e p il numero di termini del vocabolario, conterrà il numero di volte in cui ciascun termine è utilizzato nei diversi documenti della raccolta.
In Tabella 4.12 sono descritte le caratteristiche delle 13 tabelle lessicali ottenute, in termini di numero di documenti, numero di termini diversi in ciascuna raccolta e numero di classi in cui le diverse raccolte sono state categorizzate.

Tabella 4.12: Descrizione dei dataset testuali

Dataset	**N.Documenti**	**Termini**	**N.Classi**
fbis	2463	2000	17
k1a	2340	21839	20
k1b	2340	21839	6
la1	3204	21604	6
la2	3075	31472	6
re0	1504	2886	13
re1	1657	3578	25
tr11	414	6429	9
tr12	313	5804	8
tr23	204	5832	6
tr41	878	7454	10
tr45	690	8261	10
wap	1560	8460	20

Su ciascuno dei dataset così ottenuti è stato eseguito lo *Sferical-Kmeans* un numero di volte compreso tra 2 e $2g$, dove g è il numero "vero" di gruppi nello specifico dataset. Per ognuna delle partizioni è stato calcolato il valore assunto da ciascuno dei 15 indici interni di validazione.

4.2.2 Risultati

I risultati ottenuti dall'analisi comparativa degli indici interni di validazione, condotta sulle matrici lessicali ottenute dai diversi *corpora*, conferma le considerazioni fatte precedentemente nell'analisi sui dataset quantitativi. Nella tabella 4.13 è possibile osservare ciascuno dei valori degli indici di validazione calcolati per le diverse partizioni in termini di rango. Come detto, un valore pari a 1 indica in particolare che l'indice ha individuato come miglior soluzione possibile tra tutte le $(2g-1)$ partizioni quella vera, mentre un valore più elevato (in ordine crescente) implica un maggior allontanamento dalla vera classificazione dei dati. Si rimanda comunque all'Appendice A per i risultati completi.

Tabella 4.13: Sferical-Kmeans Dataset testuali

		GRAPH-BASED								**PROTOTYPE-BASED**						
Dataset	**N.Classi**	C	D	G	G_p	MCR	PB	S	T	CH	DB	PBM	RL	RT	WG	XB
fbis	17	18	6	17	18	21	20	15	19	16	7	18	15	22	30	17
k1a	20	24	37	23	17	24	30	32	30	23	30	37	20	24	34	38
k1b	6	2	2	2	7	1	2	7	4	5	3	6	5	2	3	3
la1	6	8	2	9	6	9	8	11	8	5	8	5	5	8	10	6
la2	6	3	1	1	7	1	3	5	3	5	8	5	5	4	5	2
re0	13	17	1	17	15	17	19	14	19	12	12	12	14	8	9	1
re1	25	12	30	6	26	6	7	24	4	22	28	23	15	28	25	37
tr11	9	8	1	11	9	2	3	17	9	6	7	8	8	13	16	10
tr12	8	9	2	14	9	8	6	13	14	3	13	1	1	4	14	10
tr23	6	2	3	3	8	2	2	5	4	2	8	1	1	3	3	4
tr41	10	14	10	14	10	14	14	14	18	18	10	14	14	18	17	15
tr45	10	10	1	2	11	9	6	7	3	8	6	14	6	9	1	7
wap	20	21	32	17	23	17	18	15	20	19	29	22	19	16	15	36

É semplice intuire guardando ai ranghi come solo in pochissimi casi gli indicatori forniscono una risposta adeguata in termini di numero corretto di gruppi. In realtà non si poteva che prevedere una situazione del genere, risultati per lo più discordanti e di conseguenza poco affidabili.

Da un lato questi risultati possono essere il frutto delle criticità evidenziate nell'analisi comparativa preliminare, in tal senso gli indicatori non forniscono una soluzione che si avvicini al numero "vero di gruppi" perché il metodo non riesce a catturare la struttura sottostante i dati; dall'altro, poiché le diverse misure interne di validazione misurano la coesione e/o la separazione, seguendo un approccio *graph-based* o *prototype-based*, in termini di distanza concetto che può diventare molto labile all'aumentare delle dimensioni del dataset di partenza.

4.3 Discussione

L'analisi comparativa sugli indici di validazione presentata in questo lavoro segue le linee guida consolidate dalla letteratura sulla *Cluster Validation* sviluppata nell'arco degli ultimi $15 - 20$ anni.
Alle luce dei risultati fin qui presentati più che risposte convincenti relativamente alle performance di questi indicatori sembrerebbero essere sorte ulteriori domande. Il problema è quindi duplice: da un lato gli indicatori mancano di considerare altri aspetti, oltre alla coesione e alla separazione tra i gruppi, in quanto dovrebbero tener conto anche di altre caratteristiche dei gruppi individuati quali ad esempio la densità, o la forma dei gruppi. Dall'altro questo modo di procedere per valutare le performance degli indicatori, seppur coadiuvato da anni di studi sul tema, presenta notevoli criticità. È bene quindi sviluppare alcune considerazioni che tengano conto della strategia d'analisi adottata, soffermandosi in prima istanza sull'analisi comparativa dei dataset quantitativi e considerando in particolare i risultati ottenuti a partire dai dataset simulati.
In generale i dati simulati consentono di osservare le performance di determi-

nati strumenti avendo sotto sottocontrollo la struttura dei dati e conoscendo quindi quella che si definisce la partizione "vera" dei dati. Gli esempi presentati per i dataset simulati a due dimensioni hanno l'intento di evidenziare il legame tra questi tre fattori: la struttura dei dati, il criterio di raggruppamento e conseguentemente la validazione dei risultati attraverso gli indici.
È stato possibile notare che solo alcuni indici riescono ad individuare il numero corretto di gruppi, misurando la coesione e la separazione dei gruppi ottenuti, ma individuano la partizione "migliore" secondo questi due criteri, senza tener conto di altre valutazioni. La prima domanda da porsi allora è se gli indicatori esaminati in questo lavoro, e comunemente utilizzati in letteratura, tengono conto di tutti gli aspetti, considerando che diversi metodi di *Clustering* ottimizzando un dato criterio presuppongono a monte una diversa definizione di gruppo.
Fissando di volta in volta il criterio di raggruppamento è possibile osservare come i dati vengano partizionati dai diversi algoritmi. Quando è stata indicata come miglior soluzione possibile tra quelle ottenute la partizione cui corrisponde il numero corretto di gruppi, è stato possibile notare come questa partizione non corrisponde spesso alla partizione "vera" dei dati.
Nello specifico, dato un dataset per cui è noto che il metodo di *Clustering* (*kmeans*) non è in grado di individuare correttamente i gruppi, il risultato ottenuto dall'indicatore seppur corretto in termini di numero di gruppi non lo è in termini di composizione dei gruppi.
Solo se il metodo partiziona esattamente i dati secondo le "vere" etichette di classe, per cui la partizione ottenuta riflette la "naturale struttura" dei dati, allora, e solo in questo caso, è possibile affermare se alcuni indicatori consentono di ottenere risultati migliori di altri. Di conseguenza per fornirne una comparazione adeguata si dovrebbero confrontare le performance

dei diversi indici solo per quei dataset in cui l'algoritmo riesce a catturare la struttura sottostante i dati.
Queste considerazioni sembrano banali su dataset simulati e per algoritmi tradizionali, non lo sono più così tanto se si pensa a quanto questi metodi siano applicati su dati reali, di cui non si hanno a disposizione altre informazioni se non i dati stessi.
Dall'analisi delle performance degli indicatori, per ciò che concerne l'ambito specifico del *Clustering* di documenti, è emerso infatti che seppure si valutano le risposte fornite dai singoli indici interni queste risultano poco affidabili, probabilmente perché misurando la validità della soluzione in termini di distanza, questo concetto perde di significatività in spazi altamente dimensionali.
L'individuazione di *cluster* in spazi altamente dimensionali risulta, infatti, alquanto problematica a causa della cosiddetta *curse of dimensionality* [5]. L'aumento del numero delle caratteristiche, e di conseguenza delle dimensioni, comporta una serie di problemi nel calcolo delle misure di distanza e di similarità, in quanto i dati appaiono dispersi nello spazio. Molti studi [8] [36] hanno dimostrato che al crescere del numero di dimensioni di un dataset i concetti di similarità e di distanza, essenziali per l'analisi dei *cluster*, diventano poco significativi, in quanto la differenza fra la distanza di un punto dal punto più lontano e quella dal punto più vicino tende a zero. Inoltre per la valutazione degli indicatori in questo contesto specifico il problema riguarda anche la veridicità effettiva delle etichette con cui sono categorizzate le raccolte documentarie comunemente utilizzate in letteratura. Queste sono infatti il frutto di un processo di categorizzazione umana ed in quanto tali non riflettono necessariamente la partizione "migliore" dei documenti della raccolta, ma bensì una tra le possibili. Sebbene, quindi,

l'utilizzo di questi indicatori è principalmente finalizzato ad individuare il numero di gruppi per cui si ottiene la partizione con gruppi più coesi e meglio separati, non è possibile asserire se siano in grado di dare una misura complessiva della qualità del risultato di *Clustering*.

Conclusioni

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.
Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

L'affermazione riportata nel libro di Jain e Dubes *"Algorithms for clustering data"* [42] è quanto mai vera ed attuale, nonostante da allora siano passati ben 27 anni. Nel tempo le tecniche di *Cluster analysis* si sono sviluppate in maniera esponenziale da un lato per far fronte ad un bisogno informativo proveniente dai più svariati campi di ricerca; dall'altro per essere in grado di gestire volumi di dati in continua crescita.
La valutazione del prodotto di queste tecniche è una questione che per troppo tempo è stata messa da parte. Forse perché la più complessa e più "scomoda" da affrontare, essendo più semplice proporre una nuova soluzione per la classificazione automatica, piuttosto che valutare la capacità dei metodi esistenti di rispondere all'esigenza per cui sono stati ideati: individuare, se esistono, dei raggruppamenti nascosti nei dati.
In questo lavoro si è voluto porre l'accento su una parte di quanto è stato

prodotto in tema di validazione negli ultimi trent'anni, per valutare quale tra i metodi esistenti fosse più idoneo per lo specifico campo applicativo del *Clustering* di documenti.
Le domande di ricerca che ci si è posti hanno sollevato altre domande. Si sono infatti volute comparare, in un'ampia varietà di configurazioni, le performance di diversi indici interni di validazione su dati di tipo quantativo, perseguendo l'obiettivo di individuare quali di essi dessero delle risposte più adeguate, per poi utilizzarli in un campo più complesso come quello del *Clustering* di documenti.
Il lavoro è stato sviluppato seguendo le linee guida della *Cluster validation*. Nei contributi scientifici in cui si comparano le performance degli indici interni spesso si conclude che nessuno di essi mostra migliori prestazioni degli altri. Concordemente infatti gli indici interni per la validazione sembrano il più delle volte fornire della soluzioni differenti evidenziando migliori performance di alcuni indicatori in alcuni a casi piuttosto che in altri.
Tuttavia ad un più attento esame ci si rende conto che non è questa la conclusione a cui giungere, in realtà il confronto tra gli indicatori non è valutabile, in quanto la metodologia di comparazione adottata è scorretta.
Nel lavoro è infatti evidenziato come, sebbene in alcuni casi gli indici individuano la partizione migliore in termini di numero di gruppi, questa sia effettivamente diversa dalla naturale struttura dei dati, poiché a monte il metodo di *Clustering* scelto non riesce ad individuarla. Pertanto gli indici dovrebbero essere confrontati sulla base di una strategia che tenga conto della struttura dei dati unitamente al metodo che sia in grado di individuarla.
Dall'analisi delle performance degli indicatori, per ciò che concerne l'ambito specifico del *Clustering* di documenti, è emerso che seppure si valutano

le risposte fornite dai singoli indici interni queste risultano poco affidabili, probabilmente perché misurando la validità della soluzione in termini di distanza, questo concetto perde di significatività in spazi altamente dimensionali.
Inoltre per la valutazione degli indicatori in questo contesto specifico il problema riguarda anche la veridicità effettiva delle etichette con cui sono categorizzate le raccolte documentarie comunemente utilizzate in letteratura. Queste sono infatti il frutto di un processo di categorizzazione umana ed in quanto tali non riflettono necessariamente la partizione "migliore" dei documenti della raccolta, ma bensì una tra le possibili.
Le criticità evidenziate in tema di *Cluster validation* in generale e nel contesto specifico del *Text clustering* portano alla conclusione che gli strumenti per la validazione della qualità di una soluzione di *Clustering* ad oggi disponibili, analizzati in questo lavoro, necessitano di essere migliorati per evolversi di pari passo con le nuove soluzioni per la classificazione automatica, al fine di fornire una misura dell'affidabilità del risultato ottenuto.
In tal senso, non si può che riferirsi ancora una volta all'affermazione di Jain a Dubes. . . senza uno sforzo in questa direzione, la *Cluster analysis* resterà un'arte "oscura" accessibile solo a coloro che davvero credono di avere esperienza e coraggio.

Appendice A

Tabelle dei risultati

Tabella A.1: Aggregation

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.13	704.79	0.90	0.02	-0.72	0.43	0.54	171.89	-4.98	0.25	0.45	0.47	-0.50	0.56	276.16
kmeans.3	0.06	1053.64	0.68	0.02	-0.89	0.43	0.40	225.18	-5.89	0.17	0.49	0.52	-0.60	0.63	209.06
kmeans.4	0.08	943.82	0.62	0.02	-0.85	0.35	0.38	192.32	-5.22	0.41	0.44	0.45	-0.52	0.57	177.47
kmeans.5	0.04	1326.07	0.88	0.03	-0.93	0.32	0.30	243.78	-5.36	0.25	0.42	0.51	-0.53	0.61	105.27
kmeans.6	0.05	1177.01	0.89	0.03	-0.92	0.29	0.30	195.47	-5.05	0.28	0.38	0.45	-0.50	0.57	93.65
kmeans.7	0.03	1363.92	0.67	0.03	-0.95	0.27	0.26	196.54	-5.02	0.20	0.36	0.50	-0.50	0.59	69.56
kmeans.8	0.03	1400.43	0.78	0.03	-0.95	0.23	0.24	191.94	-4.67	0.34	0.34	0.47	-0.46	0.56	79.11
kmeans.9	0.04	1135.07	0.61	0.03	-0.93	0.25	0.27	161.55	-4.84	0.68	0.32	0.46	-0.48	0.59	64.66
kmeans.10	0.03	1355.76	0.73	0.02	-0.94	0.19	0.23	207.81	-4.31	0.35	0.31	0.48	-0.42	0.57	98.11
kmeans.11	0.04	1157.29	0.81	0.03	-0.93	0.20	0.25	143.57	-4.37	0.94	0.29	0.41	-0.43	0.53	67.53
kmeans.12	0.03	1341.00	0.39	0.03	-0.94	0.17	0.22	161.69	-4.04	0.42	0.28	0.48	-0.39	0.59	48.12
kmeans.13	0.03	1408.66	0.39	0.03	-0.94	0.16	0.21	154.56	-3.92	0.37	0.27	0.48	-0.38	0.59	44.16
kmeans.14	0.04	1199.25	0.76	0.02	-0.93	0.16	0.22	128.85	-3.88	0.53	0.26	0.46	-0.38	0.59	123.87

Tabella A.2: Wine

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.50	23.53	1.88	0.00	-0.05	0.26	0.98	14996.62	-4.20	1.73	0.38	0.19	-0.03	0.00	5438.43
kmeans.3	0.18	199.30	0.59	0.00	-0.51	0.34	0.45	143824.73	-111.82	2.53	0.39	0.20	-0.34	0.08	1881.07
kmeans.4	0.18	154.98	4.52	0.00	-0.49	0.30	0.46	99220.57	-99.98	11.12	0.35	0.07	-0.31	0.13	2223.38
kmeans.5	0.19	107.40	26.08	0.00	-0.46	0.27	0.48	59949.45	-90.84	1784.64	0.32	-0.02	-0.28	0.00	1770.11
kmeans.6	0.19	97.22	2.02	0.00	-0.42	0.22	0.50	51256.12	-76.80	180.86	0.30	-0.05	-0.23	0.17	1611.38

Tabella A.3: Iris

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.02	513.30	0.41	0.08	-0.96	0.49	0.32	19.86	-1.36	0.07	0.56	0.70	-0.67	0.77	7.81
kmeans.3	0.03	560.40	0.66	0.10	-0.91	0.43	0.27	25.07	-1.17	0.16	0.50	0.56	-0.61	0.67	7.52
kmeans.4	0.03	529.02	0.31	0.10	-0.92	0.37	0.25	20.65	-1.05	0.23	0.44	0.47	-0.57	0.62	6.37
kmeans.5	0.02	494.09	0.61	0.08	-0.93	0.35	0.23	19.03	-1.00	0.29	0.40	0.44	-0.56	0.61	7.76
kmeans.6	0.04	383.14	0.73	0.10	-0.86	0.29	0.25	12.08	-0.87	0.79	0.38	0.35	-0.48	0.51	6.35

Tabella A.4: Glass

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.07	135.31	1.06	0.09	-0.89	0.44	0.36	6.54	-1.50	0.28	0.30	0.43	-0.60	0.65	3.31
kmeans.3	0.05	96.73	1.00	0.09	-0.93	0.46	0.33	7.15	-1.59	0.23	0.31	0.35	-0.64	0.66	2.83
kmeans.4	0.02	122.20	0.89	0.17	-0.97	0.48	0.30	9.73	-1.69	0.16	0.33	0.38	-0.68	0.69	1.65
kmeans.5	0.04	122.94	0.86	0.03	-0.91	0.45	0.28	8.61	-1.40	0.59	0.32	0.32	-0.63	0.57	54.88
kmeans.6	0.03	124.51	0.94	0.05	-0.93	0.46	0.27	9.99	-1.43	0.48	0.31	0.32	-0.64	0.58	17.76
kmeans.7	0.07	112.34	0.94	0.03	-0.79	0.31	0.30	8.22	-1.01	1.96	0.29	0.26	-0.47	0.44	56.48
kmeans.8	0.09	92.23	1.00	0.03	-0.72	0.25	0.34	5.61	-0.83	2.33	0.27	0.32	-0.39	0.41	37.11
kmeans.9	0.06	114.62	0.89	0.03	-0.81	0.31	0.28	7.05	-1.02	1.53	0.27	0.38	-0.47	0.47	43.92
kmeans.10	0.05	111.17	1.40	0.03	-0.82	0.31	0.28	5.97	-1.02	1.41	0.26	0.36	-0.48	0.46	40.71
kmeans.11	0.06	98.45	0.85	0.04	-0.78	0.21	0.29	5.35	-0.78	2.01	0.25	0.32	-0.38	0.44	27.43
kmeans.12	0.05	101.80	1.06	0.03	-0.82	0.24	0.26	4.78	-0.88	1.66	0.24	0.31	-0.42	0.43	36.73
kmeans.13	0.05	102.64	0.91	0.03	-0.81	0.23	0.27	6.15	-0.86	1.53	0.24		-0.41	0.44	33.72
kmeans.14	0.06	94.54	1.04	0.04	-0.79	0.20	0.28	4.13	-0.76	1.61	0.22		-0.37	0.43	22.46

Tabella A.5: Yeast

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.266	427.558	1.602	0.023	-0.491	0.365	0.710	0.028	-0.064	0.741	0.260	0.234	-0.344	0.386	64.576
kmeans.3	0.226	301.512	0.626	0.023	-0.534	0.378	0.673	0.090	-0.074	0.674	0.291	0.342	-0.375	0.394	59.130
kmeans.4	0.193	350.570	1.229	0.028	-0.557	0.327	0.663	0.028	-0.064	1.219	0.253	0.210	-0.361	0.328	44.219
kmeans.5	0.164	341.699	0.580	0.028	-0.594	0.331	0.629	0.050	-0.071	1.078	0.274	0.277	-0.383	0.335	39.312
kmeans.6	0.154	305.466	1.119	0.022	-0.630	0.255	0.615	0.017	-0.061	1.379	0.241	0.191	-0.352	0.315	58.457
kmeans.7	0.137	278.825	1.071	0.028	-0.661	0.228	0.590	0.016	-0.061	1.270	0.229	0.166	-0.346	0.310	35.468
kmeans.8	0.130	257.007	1.460	0.029	-0.676	0.207	0.578	0.013	-0.059	1.161	0.218	0.155	-0.336	0.296	37.499
kmeans.9	0.114	295.152	1.725	0.024	-0.698	0.220	0.557	0.025	-0.064	1.204	0.259	0.228	-0.355	0.319	45.703
kmeans.10	0.109	278.035	1.349	0.024	-0.712	0.189	0.544	0.018	-0.060	1.003	0.226	0.201	-0.335	0.304	44.063
kmeans.11	0.106	261.139	1.496	0.024	-0.720	0.179	0.538	0.016	-0.059	1.299	0.218	0.186	-0.329	0.288	42.868
kmeans.12	0.107	249.255	1.309	0.030	-0.720	0.158	0.535	0.014	-0.055	1.529	0.210	0.182	-0.309	0.285	41.523
kmeans.13	0.100	236.609	1.665	0.027	-0.737	0.167	0.526	0.012	-0.058	1.258	0.204	0.178	-0.323	0.286	56.792
kmeans.14	0.100	226.047	1.947	0.026	-0.739	0.143	0.520	0.011	-0.053	1.362	0.197	0.171	-0.299	0.278	34.680
kmeans.15	0.099	227.423	1.246	0.036	-0.738	0.144	0.519	0.013	-0.054	1.632	0.210	0.205	-0.300	0.299	26.269
kmeans.16	0.097	215.036	1.262	0.030	-0.746	0.117	0.510	0.009	-0.049	1.293	0.187	0.164	-0.273	0.283	37.178
kmeans.17	0.093	211.754	1.596	0.032	-0.759	0.132	0.505	0.014	-0.053	1.064	0.185	0.218	-0.294	0.299	35.916
kmeans.18	0.087	206.752	1.311	0.032	-0.777	0.121	0.492	0.013	-0.051	1.080	0.182	0.201	-0.287	0.290	34.986
kmeans.19	0.092	197.764	1.461	0.032	-0.760	0.110	0.499	0.012	-0.048	1.318	0.175	0.188	-0.269	0.275	34.656
kmeans.20	0.090	193.720	1.201	0.034	-0.769	0.105	0.492	0.011	-0.048	1.403	0.173	0.193	-0.265	0.284	29.597

Tabella A.6: R15

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.282	275.143	1.270	0.024	-0.401	0.350	0.686	9.754	-1.080	0.523	0.397	0.308	-0.283	0.455	135.714
kmeans.3	0.223	329.095	0.997	0.006	-0.519	0.346	0.591	7.106	-1.313	0.298	0.418	0.340	-0.350	0.446	2266.174
kmeans.4	0.128	361.923	0.823	0.040	-0.716	0.375	0.463	19.241	-1.746	0.273	0.402	0.445	-0.473	0.500	63.726
kmeans.5	0.020	456.548	0.809	0.403	-0.951	0.438	0.348	20.683	-2.286	0.158	0.388	0.579	-0.637	0.567	0.686
kmeans.6	0.015	486.413	0.415	0.381	-0.963	0.435	0.338	21.044	-2.299	0.129	0.366	0.681	-0.641	0.613	0.548
kmeans.7	0.011	556.543	0.591	0.393	-0.974	0.432	0.329	22.695	-2.308	0.108	0.348	0.771	-0.644	0.673	0.421
kmeans.8	0.066	530.234	0.591	0.010	-0.858	0.231	0.283	19.048	-1.625	0.444	0.328	0.590	-0.428	0.570	655.412
kmeans.9	0.057	565.563	0.338	0.009	-0.893	0.281	0.291	19.093	-1.807	9.415	0.313	0.597	-0.487	0.583	623.597
kmeans.10	0.043	791.679	0.773	0.016	-0.904	0.186	0.218	24.828	-1.538	0.440	0.304	0.658	-0.399	0.640	131.036
kmeans.11	0.075	431.181	0.703	0.010	-0.829	0.173	0.270	12.807	-1.400	7.573	0.283	0.514	-0.361	0.572	442.651
kmeans.12	0.017	1852.301	0.221	0.026	-0.968	0.177	0.164	41.042	-1.576	1.948	0.285	0.646	-0.411	0.681	82.475
kmeans.13	0.016	2103.406	0.662	0.051	-0.972	0.157	0.145	42.702	-1.502	1.574	0.274	0.649	-0.388	0.687	21.169
kmeans.14	0.013	2455.664	0.500	0.030	-0.978	0.142	0.128	47.594	-1.441	1.427	0.265	0.650	-0.370	0.723	49.146
kmeans.15	0.014	2455.087	0.252	0.025	-0.978	0.139	0.126	45.074	-1.429	1.773	0.256	0.601	-0.367	0.705	61.638
kmeans.16	0.002	4659.596	0.157	0.062	-0.997	0.118	0.087	73.732	-1.353	0.633	0.249	0.689	-0.342	0.792	16.303
kmeans.17	0.004	4473.090	0.386	0.070	-0.995	0.114	0.087	67.926	-1.327	0.731	0.242	0.648	-0.336	0.769	12.186
kmeans.18	0.016	2022.800	0.836	0.015	-0.975	0.137	0.129	33.339	-1.414	2.208	0.234	0.558	-0.363	0.707	193.396
kmeans.19	0.005	4212.976	0.372	0.054	-0.993	0.106	0.085	56.149	-1.279	0.612	0.229	0.568	-0.323	0.732	19.672
kmeans.20	0.005	4089.944	0.672	0.038	-0.993	0.105	0.084	54.239	-1.273	0.665	0.223	0.563	-0.322	0.734	37.912
kmeans.21	0.016	2405.701	0.142	0.022	-0.975	0.112	0.112	36.013	-1.287	1.358	0.217	0.482	-0.329	0.650	56.302
kmeans.22	0.007	3860.634	0.375	0.037	-0.988	0.094	0.084	46.745	-1.205	0.827	0.212	0.455	-0.305	0.640	38.209
kmeans.23	0.006	3794.650	0.696	0.038	-0.991	0.098	0.083	44.873	-1.233	0.899	0.208	0.538	-0.312	0.733	35.127
kmeans.24	0.007	3590.394	0.302	0.027	-0.989	0.098	0.084	43.492	-1.231	0.648	0.203		-0.311	0.706	71.722
kmeans.25	0.009	3571.613	0.411	0.045	-0.985	0.089	0.084	38.729	-1.165	0.891	0.199	0.447	-0.295	0.644	24.542
kmeans.26	0.007	3694.619	0.401	0.041	-0.988	0.088	0.080	38.756	-1.165	0.555	0.196	0.455	-0.294	0.648	29.088
kmeans.27	0.009	3415.011	0.852	0.034	-0.984	0.086	0.083	36.068	-1.144	0.753	0.192	0.424	-0.289	0.629	40.967
kmeans.28	0.076	353.452	0.589	0.009	-0.825	0.091	0.226	8.272	-1.037	6.854	0.184	0.388	-0.260	0.565	305.229
kmeans.29	0.009	3508.304	0.652	0.026	-0.984	0.078	0.078	33.620	-1.090	1.042	0.185	0.424	-0.275	0.632	66.718
kmeans.30	0.010	3360.361	0.489	0.037	-0.982	0.077	0.080	33.069	-1.081	1.272	0.182	0.418	-0.273	0.626	31.388

Tabella A.7: Flame

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.189	157.206	1.115	0.033	-0.601	0.4	0.606	15.424	-1.458	0.378	0.442	0.379	-0.425	0.502	71.319
kmeans.3	0.116	202.721	0.792	0.04	-0.755	0.395	0.495	14.42	-1.724	0.204	0.457	0.418	-0.507	0.543	43.688
kmeans.4	0.067	258.164	0.634	0.048	-0.865	0.362	0.411	23.423	-1.829	0.164	0.437	0.448	-0.539	0.57	25.238
kmeans.5	0.073	238.466	0.653	0.046	-0.857	0.316	0.396	21.026	-1.706	0.342	0.4	0.401	-0.5	0.535	23.409
kmeans.6	0.065	242.326	0.839	0.054	-0.876	0.264	0.361	19.161	-1.6	0.326	0.373	0.392	-0.465	0.523	13.994

Tabella A.8: Jain toy

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.119	503.475	0.783	0.019	-0.735	0.434	0.476	196.55	-5.175	0.184	0.518	0.493	-0.519	0.602	267.6
kmeans.3	0.08	580.555	0.734	0.032	-0.827	0.407	0.385	280.951	-5.331	0.177	0.496	0.492	-0.552	0.598	52.176
kmeans.4	0.062	594.68	0.791	0.018	-0.869	0.354	0.328	310.315	-5.14	0.191	0.451	0.458	-0.535	0.573	131.796
kmeans.5	0.05	657.419	0.733	0.015	-0.895	0.309	0.288	330.793	-4.911	0.172	0.415	0.463	-0.511	0.575	186.256
kmeans.6	0.038	713.622	0.562	0.017	-0.92	0.27	0.25	323.524	-4.698	0.16	0.385	0.465	-0.488	0.587	141.5

Tabella A.9: S2

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.165	3912.218	1.037	0.008	-0.649	0.412	0.577	8.45E+10	-1.10E+05	0.318	0.459	0.409	-0.459	0.521	1023.860
kmeans.3	0.147	3470.736	0.880	0.003	-0.680	0.385	0.530	5.72E+10	-1.11E+05	0.296	0.440	0.378	-0.461	0.489	8386.875
kmeans.4	0.064	5521.916	0.813	0.006	-0.858	0.351	0.402	9.09E+10	-1.25E+05	0.194	0.438	0.447	-0.528	0.543	2108.857
kmeans.5	0.057	5395.387	0.806	0.007	-0.877	0.311	0.371	8.68E+10	-1.21E+05	0.253	0.403	0.448	-0.504	0.529	1496.492
kmeans.6	0.055	5262.282	0.773	0.003	-0.884	0.270	0.345	7.22E+10	-1.15E+05	0.250	0.374	0.439	-0.474	0.513	6313.855
kmeans.7	0.051	5210.081	0.747	0.003	-0.893	0.244	0.323	7.86E+10	-1.11E+05	0.320	0.350	0.448	-0.453	0.523	7854.150
kmeans.8	0.051	5243.030	0.579	0.007	-0.898	0.231	0.314	7.71E+10	-1.09E+05	0.367	0.332	0.474	-0.443	0.535	1154.939
kmeans.9	0.046	5639.974	0.675	0.007	-0.912	0.205	0.287	8.10E+10	-1.05E+05	0.309	0.316	0.488	-0.422	0.554	960.740
kmeans.10	0.028	7054.668	0.721	0.007	-0.954	0.189	0.242	1.05E+11	-1.05E+05	0.159	0.304	0.551	-0.419	0.614	909.475
kmeans.11	0.036	6300.266	0.477	0.005	-0.937	0.172	0.241	8.20E+10	-1.00E+05	0.227	0.290	0.534	-0.395	0.605	1401.090
kmeans.12	0.028	6946.272	0.463	0.004	-0.957	0.171	0.227	9.18E+10	-1.01E+05	1.594	0.280	0.512	-0.400	0.615	3122.482
kmeans.13	0.027	7514.257	0.808	0.009	-0.960	0.159	0.213	1.06E+11	-9.88E+04	0.891	0.270	0.552	-0.387	0.650	551.738
kmeans.14	0.011	11766.987	0.381	0.008	-0.988	0.136	0.161	1.60E+11	-9.57E+04	0.123	0.263	0.623	-0.365	0.717	490.531
kmeans.15	0.015	10550.377	0.343	0.003	-0.983	0.132	0.165	1.23E+11	-9.41E+04	0.614	0.254	0.561	-0.359	0.684	3116.385
kmeans.16	0.017	9946.422	0.301	0.007	-0.981	0.127	0.163	1.16E+11	-9.21E+04	0.533	0.246	0.535	-0.351	0.674	502.344
kmeans.17	0.016	10021.784	0.492	0.004	-0.983	0.127	0.161	1.19E+11	-9.23E+04	1.034	0.239	0.521	-0.352	0.673	2266.417
kmeans.18	0.033	6790.969	0.592	0.004	-0.957	0.132	0.196	7.05E+10	-9.09E+04	1.281	0.231	0.441	-0.352	0.604	1425.265
kmeans.19	0.017	9402.641	0.408	0.012	-0.982	0.122	0.157	9.94E+10	-9.06E+04	0.943	0.226	0.485	-0.344	0.655	221.414
kmeans.20	0.012	11221.492	0.720	0.005	-0.988	0.109	0.136	1.21E+11	-8.71E+04	0.608	0.221	0.486	-0.327	0.666	956.547
kmeans.21	0.022	8192.697	0.846	0.006	-0.976	0.118	0.162	7.66E+10	-8.85E+04	0.836	0.215	0.451	-0.337	0.634	814.412
kmeans.22	0.012	10690.296	0.782	0.011	-0.988	0.104	0.131	9.37E+10	-8.55E+04	0.551	0.211	0.452	-0.320	0.648	251.848
kmeans.23	0.013	10727.678	0.892	0.008	-0.987	0.098	0.127	9.44E+10	-8.30E+04	0.699	0.206	0.426	-0.309	0.623	378.056
kmeans.24	0.014	10128.687	0.780	0.008	-0.987	0.100	0.130	8.21E+10	-8.39E+04	0.880	0.202	0.439	-0.313	0.640	400.678
kmeans.25	0.015	10149.323	0.538	0.009	-0.986	0.093	0.125	7.80E+10	-8.09E+04	0.663	0.198	0.391	-0.301	0.595	370.691
kmeans.26	0.014	10121.034	0.844	0.006	-0.987	0.092	0.123	7.46E+10	-8.08E+04	0.538	0.194	0.401	-0.300	0.601	807.970
kmeans.27	0.013	10053.921	0.275	0.009	-0.987	0.091	0.121	8.06E+10	-8.05E+04	0.571	0.191	0.406	-0.299	0.611	379.333
kmeans.28	0.016	9643.541	0.662	0.007	-0.984	0.088	0.124	6.85E+10	-7.89E+04	0.669	0.187	0.382	-0.293	0.581	650.034
kmeans.29	0.016	9466.537	0.586	0.005	-0.985	0.086	0.121	6.64E+10	-7.80E+04	0.857	0.184	0.378	-0.290	0.581	1511.957
kmeans.30	0.013	9762.790	0.749	0.006	-0.987	0.085	0.116	6.18E+10	-7.80E+04	0.499	0.181	0.386	-0.289	0.584	612.992

Tabella A.10: Pathbased

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.102	371.055	0.752	0.037	-0.79	0.447	0.479	154.837	-4.464	0.201	0.457	0.515	-0.559	0.652	41.707
kmeans.3	0.049	359.105	0.679	0.034	-0.897	0.457	0.398	107.274	-4.954	0.123	0.471	0.517	-0.623	0.657	51.261
kmeans.4	0.076	318.754	1.089	0.028	-0.843	0.395	0.393	124.066	-4.389	0.428	0.43	0.417	-0.552	0.571	75.393
kmeans.5	0.069	297.471	0.782	0.036	-0.857	0.368	0.37	110.077	-4.295	0.447	0.393	0.405	-0.54	0.561	43.843
kmeans.6	0.057	310.265	0.768	0.023	-0.884	0.31	0.325	106.333	-4.039	0.34	0.371	0.38	-0.507	0.513	132.746

Tabella A.11: Transfusion

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.149	438.829	0.98	0.015	-0.721	0.37	0.494	3.572	-0.837	0.266	0.43	0.382	-0.473	0.57	93.41
kmeans.3	0.133	412.473	0.822	0.008	-0.672	0.395	0.499	2.822	-0.7	0.616	0.419	0.319	-0.462	0.475	281.605
kmeans.4	0.106	474.908	0.958	0.005	-0.714	0.378	0.462	10.744	-0.714	0.411	0.405	0.361	-0.474	0.499	814.326
kmeans.5	0.086	490.122	0.851	0.005	-0.764	0.349	0.427	13.467	-0.702	0.32	0.381	0.403	-0.48	0.487	652.37
kmeans.6	0.066	486.777	0.994	0.015	-0.822	0.319	0.387	11.43	-0.692	0.345	0.357	0.386	-0.486	0.494	61.621

Tabella A.12: Dim032

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.184	165.697	2.449	0.550	-0.432	0.358	0.824	15431.140	-40.299	1.518	0.258	0.180	-0.306	0.217	0.964
kmeans.3	0.140	167.347	2.142	0.519	-0.572	0.377	0.774	14708.950	-50.816	1.076	0.283	0.303	-0.396	0.283	0.844
kmeans.4	0.135	149.767	1.800	0.556	-0.577	0.359	0.754	22264.151	-53.819	0.973	0.273	0.454	-0.390	0.322	0.778
kmeans.5	0.101	169.899	1.700	0.555	-0.711	0.306	0.661	11877.944	-65.191	0.733	0.284	0.420	-0.425	0.357	0.672
kmeans.6	0.127	153.426	2.121	0.556	-0.589	0.256	0.645	11239.978	-64.201	1.225	0.266	0.510	-0.335	0.367	0.542
kmeans.7	0.090	185.149	0.567	0.635	-0.737	0.288	0.630	14131.797	-68.311	0.566	0.270	0.642	-0.424	0.488	0.454
kmeans.8	0.091	205.468	0.540	0.525	-0.727	0.236	0.564	11944.993	-72.541	0.553	0.269	0.654	-0.380	0.517	0.464
kmeans.9	0.056	254.033	0.535	0.641	-0.888	0.227	0.488	13344.732	-79.875	0.395	0.270	0.696	-0.435	0.570	0.373
kmeans.10	0.076	198.131	1.106	0.055	-0.792	0.205	0.484	9203.442	-78.318	31.942	0.253		-0.379	0.533	37.541
kmeans.11	0.087	227.652	1.798	0.037	-0.748	0.199	0.491	11495.352	-76.951	93.432	0.251	0.614	-0.357	0.600	82.225
kmeans.12	0.064	299.819	0.696	0.026	-0.813	0.196	0.450	16381.691	-81.058	177.770	0.251	0.679	-0.378	0.668	131.771
kmeans.13	0.106	169.089	0.578	0.028	-0.727	0.175	0.458	7238.977	-76.818	217.349	0.229	0.410	-0.327	0.503	135.512
kmeans.14	0.022	387.541	0.484	0.045	-0.966	0.165	0.293	16109.015	-91.619	56.563	0.244		-0.396	0.712	44.974
kmeans.15	0.053	413.643	0.697	0.043	-0.911	0.158	0.316	25973.021	-87.743	77.447	0.239	0.652	-0.371	0.730	39.961
kmeans.16	0.062	220.861	0.936	0.043	-0.917	0.170	0.357	7617.173	-85.339	121.761	0.218		-0.386	0.539	62.826
kmeans.17	0.048	280.026	0.680	0.043	-0.921	0.172	0.349	11310.997	-86.905	55.677	0.219		-0.389	0.664	48.999
kmeans.18	0.011	762.751	0.242	0.052	-0.996	0.140	0.182	35218.903	-96.685	24.386	0.226		-0.374	0.813	19.390
kmeans.19	0.058	345.805	0.658	0.031	-0.924	0.144	0.267	15270.721	-89.132	103.409	0.213		-0.357	0.632	71.194
kmeans.20	0.023	1246.488	0.282	0.029	-0.993	0.126	0.126	68168.388	-97.507	38.048	0.219		-0.353	0.810	36.300
kmeans.21	0.059	336.796	1.931	0.022	-0.968	0.159	0.308	14977.564	-87.552	274.617	0.203		-0.389	0.631	176.038
kmeans.22	0.052	532.869	1.505	0.017	-0.982	0.134	0.201	20664.413	-92.187	133.857	0.203		-0.361	0.689	197.495
kmeans.23	0.062	464.775	1.615	0.033	-0.956	0.145	0.265	25630.818	-88.845	82.030	0.199		-0.368	0.673	48.586
kmeans.24	0.035	1086.832	1.397	0.039	-0.995	0.123	0.125	57114.613	-96.272	28.805	0.200		-0.349	0.732	19.074
kmeans.25	0.058	425.295	0.464	0.033	-0.957	0.146	0.264	21352.508	-89.257	76.347	0.191		-0.369	0.713	48.860
kmeans.26	0.064	685.200	1.677	0.026	-0.953	0.118	0.145	37629.469	-92.864	87.246	0.191		-0.331	0.703	43.725
kmeans.27	0.002	16821.904	0.097	0.180	-0.997	0.103	0.040	302693.590	-96.285	2.575	0.192		-0.321	0.661	1.491
kmeans.28	0.068	633.378	0.425	0.032	-0.952	0.117	0.146	32217.982	-92.386	47.024	0.184		-0.329	0.630	28.186
kmeans.29	0.084	697.675	2.314	0.021	-0.974	0.113	0.143	33569.550	-90.607	78.934	0.181		-0.330	0.610	68.822
kmeans.30	0.002	15208.883	0.259	0.169	-0.998	0.106	0.041	251964.719	-97.781	1.548	0.182		-0.326	0.718	1.434
kmeans.31	0.002	14636.729	0.745	0.138	-0.998	0.106	0.041	232255.655	-97.506	2.554	0.179		-0.325	0.722	2.158
kmeans.32	0.002	14371.485	0.536	0.103	-0.997	0.101	0.040	220267.070	-94.994	2.843	0.177		-0.317	0.623	3.856

Tabella A.13: Wineqred

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.114	1285.054	0.942	0.011	-0.742	0.414	0.496	1816.488	-18.319	0.242	0.226	0.417	-0.512	0.565	136.111
kmeans.3	0.147	960.081	1.298	0.008	-0.652	0.398	0.511	1910.134	-15.457	0.639	0.21	0.297	-0.453	0.41	199.338
kmeans.4	0.075	1279.484	0.812	0.01	-0.832	0.364	0.393	2147.916	-16.726	0.347	0.22	0.415	-0.524	0.567	91.649
kmeans.5	0.072	1102.324	0.832	0.008	-0.838	0.335	0.378	1677.993	-16.109	0.524	0.202	0.369	-0.506	0.539	116.373
kmeans.6	0.049	1218.618	0.78	0.008	-0.892	0.313	0.334	2302.518	-16.239	0.412	0.191	0.395	-0.513	0.552	155.792
kmeans.7	0.057	1099.715	0.804	0.008	-0.869	0.268	0.332	2020.921	-14.754	0.905	0.18	0.355	-0.466	0.507	166.189
kmeans.8	0.048	1089.745	1.01	0.007	-0.889	0.246	0.309	2079.951	-14.39	0.777	0.171	0.345	-0.453	0.492	210.072
kmeans.9	0.049	1021.993	0.654	0.009	-0.885	0.216	0.301	1756.08	-13.449	0.975	0.163	0.332	-0.423	0.481	116.827
kmeans.10	0.05	950.471	0.904	0.008	-0.881	0.196	0.296	1581.478	-12.77	0.988	0.156	0.326	-0.402	0.483	140.903
kmeans.11	0.047	910.282	1.118	0.009	-0.886	0.176	0.284	1600.702	-12.201	1.034	0.149	0.298	-0.383	0.449	126.327
kmeans.12	0.045	869.324	0.533	0.01	-0.891	0.16	0.275	1340.829	-11.674	0.964	0.144	0.302	-0.366	0.459	81.091

Tabella A.14: Breast Wisconsin

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.047	1300.208	0.504	0.017	-0.926	0.441	0.242	1581515.56	-486.385	0.077	0.362	0.621	-0.627	0.78	37.061
kmeans.3	0.053	1154.93	0.631	0.007	-0.846	0.456	0.232	1816065.38	-396.29	0.234	0.323	0.484	-0.594	0.659	182.073
kmeans.4	0.043	1465.673	0.445	0.01	-0.839	0.421	0.212	3734020.64	-358.671	0.252	0.291	0.486	-0.567	0.65	96.452
kmeans.5	0.031	1621.415	0.586	0.008	-0.866	0.393	0.188	4048017.46	-338.963	0.259	0.266	0.477	-0.562	0.631	138.445
kmeans.6	0.027	1590.082	0.579	0.005	-0.87	0.321	0.168	4098584.4	-289.365	0.422	0.247	0.456	-0.51	0.596	276.348

Tabella A.15: Dim064

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.248	62.875	3.007	0.689	-0.035	0.219	0.920	17405.686	-22.281	2.476	0.170	0.184	-0.023	0.157	0.614
kmeans.3	0.156	103.416	2.964	0.730	-0.408	0.321	0.802	14354.843	-57.674	2.138	0.231	0.196	-0.276	0.183	0.556
kmeans.4	0.185	107.441	0.899	0.700	-0.305	0.313	0.831	30141.943	-50.234	1.660	0.245	0.462	-0.211	0.259	0.546
kmeans.5	0.110	125.890	1.831	0.774	-0.518	0.259	0.690	12327.587	-77.907	1.143	0.254	0.337	-0.302	0.291	0.436
kmeans.6	0.137	124.980	0.618	0.675	-0.495	0.280	0.734	23782.247	-69.871	1.137	0.250	0.605	-0.303	0.378	0.421
kmeans.7	0.066	161.576	1.541	0.774	-0.745	0.238	0.587	14891.943	-92.998	0.647	0.262	0.510	-0.389	0.421	0.368
kmeans.8	0.136	145.839	1.811	0.675	-0.468	0.244	0.696	21136.273	-75.179	0.767	0.248	0.717	-0.270	0.490	0.339
kmeans.9	0.082	142.724	1.681	0.041	-0.675	0.228	0.596	15849.283	-90.766	195.001	0.241	0.506	-0.353	0.471	105.209
kmeans.10	0.139	112.986	1.738	0.034	-0.427	0.209	0.657	12130.384	-79.514	323.441	0.222		-0.231	0.444	150.857
kmeans.11	0.115	170.551	0.792	0.037	-0.533	0.201	0.602	17914.352	-87.278	190.454	0.238	0.653	-0.273	0.582	95.188
kmeans.12	0.036	245.941	0.633	0.057	-0.908	0.195	0.440	25730.581	-108.763	89.141	0.245	0.681	-0.411	0.648	41.455
kmeans.13	0.080	123.946	2.286	0.032	-0.731	0.206	0.541	9194.493	-96.140	227.553	0.214		-0.357	0.465	153.237
kmeans.14	0.020	298.352	0.381	0.045	-0.955	0.164	0.316	20305.659	-119.915	105.597	0.237		-0.391	0.672	52.777
kmeans.15	0.040	249.656	0.714	0.042	-0.834	0.165	0.375	19512.134	-113.553	92.822	0.227		-0.354	0.682	50.081
kmeans.16	0.025	514.390	1.629	0.043	-0.937	0.137	0.213	50761.867	-126.689	52.130	0.235		-0.353	0.812	31.988
kmeans.17	0.047	223.663	0.483	0.042	-0.843	0.165	0.376	14866.141	-112.853	102.333	0.213		-0.356	0.631	49.094
kmeans.18	0.046	271.495	0.985	0.034	-0.834	0.153	0.335	25777.273	-116.044	99.983	0.213		-0.340	0.702	59.624
kmeans.19	0.030	441.462	0.829	0.036	-0.929	0.136	0.211	43351.954	-126.320	76.418	0.217		-0.348	0.766	42.515
kmeans.20	0.016	500.269	0.449	0.052	-0.995	0.140	0.196	43686.131	-128.734	39.586	0.213		-0.373	0.754	21.358
kmeans.21	0.050	373.422	0.219	0.040	-0.825	0.139	0.285	37473.795	-119.234	65.580	0.205		-0.322	0.772	29.457
kmeans.22	0.027	826.443	1.112	0.036	-0.978	0.125	0.130	99460.995	-131.913	30.495	0.207		-0.347	0.843	23.817
kmeans.23	0.028	941.281	0.366	0.041	-0.999	0.125	0.122	103226.382	-132.322	32.186	0.204		-0.353	0.786	17.814
kmeans.24	0.044	355.742	0.387	0.024	-0.948	0.134	0.212	23981.818	-124.840	115.896	0.192		-0.352	0.710	92.232
kmeans.25	0.001	35902.975	2.524	0.271	-0.999	0.112	0.029	1150831.389	-138.523	2.015	0.200		-0.335	0.850	0.624
kmeans.26	0.049	773.620	0.159	0.031	-0.995	0.122	0.127	77307.208	-129.881	76.054	0.191		-0.348	0.687	33.212
kmeans.27	0.034	793.838	0.834	0.042	-0.999	0.124	0.122	76098.136	-131.680	30.491	0.188		-0.351	0.741	17.449
kmeans.28	0.039	346.971	0.414	0.020	-0.991	0.136	0.202	22801.156	-125.984	312.525	0.179		-0.367	0.734	135.420
kmeans.29	0.065	649.576	0.190	0.028	-0.985	0.119	0.132	61014.527	-127.873	116.552	0.181		-0.341	0.725	38.107
kmeans.30	0.057	536.813	1.623	0.030	-0.953	0.120	0.138	50396.528	-128.399	63.114	0.177		-0.334	0.712	34.166
kmeans.31	0.055	551.883	0.411	0.038	-0.969	0.120	0.134	43374.800	-128.776	40.282	0.175		-0.339	0.746	21.860
kmeans.32	0.001	28507.146	0.086	0.292	-0.999	0.109	0.028	725157.845	-136.536	1.016	0.177		-0.330	0.758	0.521

Tabella A.16: Dim128

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.152	91.755	3.307	0.819	-0.273	0.318	0.866	31546.536	-57.978	2.741	0.202	0.136	-0.193	0.144	0.539
kmeans.3	0.116	100.749	2.739	0.786	-0.465	0.326	0.798	25192.227	-82.626	1.844	0.233	0.194	-0.310	0.196	0.533
kmeans.4	0.103	104.238	2.644	0.809	-0.501	0.331	0.789	47392.913	-85.802	1.799	0.240	0.412	-0.332	0.252	0.466
kmeans.5	0.117	79.835	1.866	0.035	-0.444	0.318	0.794	32877.240	-83.499	229.969	0.217		-0.294	0.239	237.484
kmeans.6	0.070	124.928	1.610	0.790	-0.629	0.239	0.638	20020.869	-119.779	0.891	0.249	0.393	-0.341	0.333	0.395
kmeans.7	0.097	100.476	0.675	0.079	-0.512	0.244	0.686	22988.555	-108.694	37.297	0.229		-0.291	0.330	38.871
kmeans.8	0.107	82.650	2.264	0.020	-0.434	0.238	0.700	19714.679	-105.164	1464.533	0.211		-0.250	0.306	635.161
kmeans.9	0.065	103.938	0.556	0.034	-0.639	0.257	0.659	22823.208	-116.483	174.428	0.221		-0.357	0.389	185.866
kmeans.10	0.085	83.631	1.076	0.034	-0.629	0.277	0.698	17673.952	-107.483	536.796	0.205		-0.367	0.353	205.312
kmeans.11	0.058	138.241	0.632	0.022	-0.778	0.242	0.600	26785.434	-127.175	1007.880	0.228		-0.406	0.506	327.984
kmeans.12	0.048	208.775	1.868	0.036	-0.747	0.179	0.461	38250.160	-148.554	92.386	0.240		-0.339	0.646	95.405
kmeans.13	0.047	148.447	1.478	0.032	-0.768	0.181	0.460	20626.983	-148.822	148.696	0.221		-0.348	0.545	153.856
kmeans.14	0.038	249.611	0.482	0.038	-0.854	0.168	0.382	40627.886	-159.616	78.759	0.233		-0.363	0.698	78.798
kmeans.15	0.055	126.863	1.442	0.035	-0.806	0.206	0.518	16766.870	-139.941	183.317	0.205		-0.384	0.493	129.061
kmeans.16	0.029	309.988	0.789	0.036	-0.885	0.146	0.278	44773.376	-172.581	90.149	0.226		-0.348	0.717	62.756
kmeans.17	0.060	190.019	0.777	0.023	-0.770	0.159	0.394	26348.561	-155.601	307.565	0.210		-0.326	0.651	174.573
kmeans.18	0.045	260.009	0.259	0.024	-0.831	0.141	0.286	35230.424	-169.740	441.862	0.212		-0.326	0.679	143.791
kmeans.19	0.081	229.119	0.155	0.023	-0.645	0.157	0.446	43939.019	-146.792	389.383	0.206		-0.282	0.724	151.966
kmeans.20	0.045	384.496	0.503	0.024	-0.840	0.141	0.283	79107.988	-170.022	287.967	0.209		-0.328	0.771	93.710
kmeans.21	0.031	373.984	1.293	0.037	-0.959	0.138	0.211	56674.784	-179.292	68.053	0.205		-0.359	0.773	43.932
kmeans.22	0.038	375.994	1.315	0.020	-0.977	0.137	0.209	53596.836	-178.658	527.012	0.200		-0.364	0.729	150.980
kmeans.23	0.027	669.194	0.470	0.036	-0.932	0.123	0.132	146493.028	-187.666	48.993	0.202		-0.332	0.821	23.452
kmeans.24	0.017	820.683	0.206	0.035	-0.999	0.127	0.119	161128.874	-190.134	39.948	0.199		-0.356	0.833	25.259
kmeans.25	0.000	58982.138	0.142	0.234	-1.000	0.113	0.022	3139559.663	-198.765	1.765	0.200		-0.336	0.858	0.680
kmeans.26	0.060	287.634	0.533	0.014	-0.821	0.138	0.288	47080.974	-167.619	567.405	0.184		-0.319	0.707	258.630
kmeans.27	0.001	53765.790	0.043	0.290	-1.000	0.112	0.022	2681607.520	-198.333	0.904	0.192		-0.335	0.818	0.330
kmeans.28	0.034	696.139	0.199	0.025	-0.998	0.124	0.121	118859.104	-187.693	100.007	0.184		-0.352	0.740	48.489
kmeans.29	0.043	547.571	0.200	0.028	-0.960	0.122	0.131	94507.253	-185.608	105.964	0.180		-0.339	0.741	39.110
kmeans.30	0.027	647.552	0.486	0.027	-0.999	0.125	0.120	104335.177	-188.724	55.235	0.178		-0.353	0.791	43.801
kmeans.31	0.001	47376.236	0.632	0.201	-0.999	0.110	0.022	2070510.994	-196.570	1.759	0.180		-0.332	0.804	0.673
kmeans.32	0.001	46246.449	0.074	0.201	-0.999	0.110	0.022	1962995.343	-196.566	1.648	0.177		-0.332	0.785	0.666

Tabella A.17: Dim256

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.127	90.873	3.315	0.832	-0.269	0.317	0.870	64580.471	-80.626	2.768	0.201	0.137	-0.190	0.143	0.543
kmeans.3	0.105	88.063	1.993	0.860	-0.350	0.294	0.894	158537.499	-61.642	1.183	0.219	0.520	-0.231	0.304	0.472
kmeans.4	0.104	98.181	0.799	0.862	-0.382	0.315	0.813	96583.008	-110.458	1.689	0.236	0.420	-0.258	0.255	0.445
kmeans.5	0.089	102.941	0.806	0.832	-0.494	0.285	0.751	68137.307	-134.607	1.416	0.240	0.439	-0.306	0.288	0.421
kmeans.6	0.083	86.255	2.416	0.022	-0.503	0.298	0.761	71825.182	-131.992	1321.027	0.223	0.318	-0.317	0.281	597.714
kmeans.7	0.076	125.078	0.698	0.881	-0.471	0.281	0.746	89670.153	-137.618	1.436	0.247	0.752	-0.291	0.431	0.318
kmeans.8	0.087	82.661	1.947	0.023	-0.480	0.313	0.783	47085.403	-123.644	847.932	0.213		-0.312	0.336	492.343
kmeans.9	0.030	170.440	1.455	0.878	-0.815	0.197	0.489	58128.826	-208.357	0.464	0.252	0.639	-0.380	0.516	0.245
kmeans.10	0.048	187.694	0.509	0.863	-0.644	0.216	0.590	79937.611	-183.686	0.454	0.250	0.803	-0.330	0.620	0.215
kmeans.11	0.058	127.925	0.726	0.028	-0.589	0.191	0.556	45811.760	-190.250	225.581	0.225		-0.288	0.505	227.089
kmeans.12	0.072	149.437	0.606	0.025	-0.491	0.195	0.603	58940.669	-177.531	246.185	0.227		-0.251	0.574	256.228
kmeans.13	0.045	184.838	0.626	0.023	-0.765	0.191	0.497	63402.682	-204.760	588.830	0.230	0.592	-0.356	0.615	266.423
kmeans.14	0.036	238.204	0.816	0.026	-0.782	0.214	0.543	108632.566	-195.993	159.808	0.232		-0.383	0.755	161.429
kmeans.15	0.032	166.106	0.697	0.018	-0.863	0.191	0.458	49973.297	-214.389	632.801	0.215		-0.390	0.616	445.043
kmeans.16	0.018	471.736	0.230	0.027	-0.916	0.136	0.209	193167.443	-260.275	80.558	0.234		-0.345	0.821	83.844
kmeans.17	0.036	146.262	0.388	0.024	-0.829	0.186	0.459	40927.353	-213.523	419.971	0.203		-0.374	0.550	236.831
kmeans.18	0.027	264.359	0.535	0.018	-0.847	0.142	0.277	70642.515	-247.748	600.868	0.213		-0.333	0.715	242.557
kmeans.19	0.043	126.324	0.637	0.018	-0.817	0.196	0.494	31013.692	-205.529	640.989	0.191		-0.379	0.586	450.801
kmeans.20	0.013	422.937	0.205	0.028	-0.994	0.141	0.199	145493.643	-262.643	129.937	0.211		-0.374	0.766	76.851
kmeans.21	0.034	352.705	0.541	0.024	-0.844	0.142	0.281	153976.251	-245.989	173.777	0.204		-0.331	0.752	99.532
kmeans.22	0.042	332.940	0.936	0.015	-0.922	0.134	0.214	102757.960	-255.470	885.905	0.199		-0.344	0.699	265.080
kmeans.23	0.047	198.728	0.694	0.021	-0.860	0.154	0.342	52099.435	-233.624	320.838	0.188		-0.350	0.654	201.766
kmeans.24	0.029	654.495	0.147	0.024	-0.966	0.124	0.125	310990.728	-271.155	117.152	0.197		-0.343	0.755	53.007
kmeans.25	0.033	317.790	0.950	0.025	-0.983	0.151	0.274	115413.270	-247.306	211.915	0.188		-0.384	0.705	101.076
kmeans.26	0.029	598.020	0.293	0.024	-0.961	0.124	0.125	252159.178	-271.050	117.686	0.190		-0.342	0.783	54.614
kmeans.27	0.027	634.896	0.338	0.019	-0.993	0.126	0.121	255877.039	-272.139	119.112	0.187		-0.353	0.751	83.770
kmeans.28	0.052	250.793	0.376	0.022	-0.801	0.137	0.288	84024.317	-241.804	186.989	0.176		-0.312	0.754	111.912
kmeans.29	0.037	532.419	0.425	0.018	-0.961	0.123	0.126	204118.178	-269.546	131.813	0.180		-0.340	0.737	92.022
kmeans.30	0.028	567.510	0.221	0.023	-0.993	0.125	0.121	206762.570	-271.831	75.801	0.177		-0.353	0.746	57.100
kmeans.31	0.000	109853.348	1.144	0.407	-1.000	0.112	0.015	8710991.593	-286.399	0.420	0.180		-0.334	0.797	0.237
kmeans.32	0.000	105972.256	1.291	0.354	-1.000	0.111	0.016	8175357.062	-286.037	0.746	0.177		-0.334	0.756	0.444

Tabella A.18: Dim512

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.054	81.888	2.099	0.917	-0.492	0.255	0.902	263010.990	-71.490	1.365	0.192	0.291	-0.287	0.244	0.460
kmeans.3	0.072	86.010	2.121	0.906	-0.436	0.313	0.894	343998.823	-87.537	1.313	0.218	0.521	-0.288	0.287	0.436
kmeans.4	0.073	94.815	2.669	0.886	-0.465	0.296	0.771	102426.557	-179.657	1.721	0.232	0.288	-0.295	0.221	0.436
kmeans.5	0.074	100.598	1.958	0.883	-0.486	0.300	0.772	205398.450	-179.308	1.746	0.238	0.528	-0.309	0.295	0.399
kmeans.6	0.068	110.035	0.737	0.877	-0.476	0.264	0.729	158006.947	-200.303	1.252	0.242	0.604	-0.285	0.366	0.353
kmeans.7	0.074	120.375	2.360	0.866	-0.441	0.275	0.753	151582.738	-188.388	1.058	0.243	0.698	-0.273	0.434	0.326
kmeans.8	0.067	135.091	0.446	0.900	-0.471	0.257	0.720	158276.110	-204.271	1.095	0.245	0.788	-0.279	0.498	0.282
kmeans.9	0.049	93.313	0.682	0.018	-0.653	0.289	0.713	96722.146	-210.091	746.860	0.217		-0.386	0.392	751.529
kmeans.10	0.062	105.453	0.673	0.019	-0.552	0.250	0.688	84180.212	-218.563	1138.053	0.220		-0.313	0.433	643.661
kmeans.11	0.039	163.671	0.787	0.030	-0.732	0.227	0.594	126286.547	-256.889	181.522	0.237		-0.375	0.598	183.748
kmeans.12	0.034	147.553	1.800	0.018	-0.719	0.187	0.499	96849.194	-287.834	497.786	0.227		-0.335	0.554	500.898
kmeans.13	0.020	253.719	0.596	0.018	-0.806	0.163	0.384	177892.270	-323.311	341.621	0.240		-0.343	0.707	337.159
kmeans.14	0.009	547.754	0.325	0.939	-0.959	0.140	0.205	502713.845	-369.984	0.114	0.250	0.915	-0.362	0.853	0.069
kmeans.15	0.036	155.254	0.610	0.020	-0.711	0.185	0.500	92059.883	-287.349	365.537	0.213		-0.331	0.587	355.034
kmeans.16	0.022	290.807	0.427	0.021	-0.884	0.158	0.336	213021.200	-335.531	216.519	0.225		-0.362	0.748	210.862
kmeans.17	0.024	140.970	0.569	0.019	-0.895	0.171	0.384	59585.485	-322.592	393.301	0.202		-0.380	0.588	396.403
kmeans.18	0.022	256.171	0.531	0.021	-0.922	0.149	0.278	140645.708	-349.987	216.360	0.213		-0.362	0.692	210.143
kmeans.19	0.025	241.410	0.506	0.018	-0.887	0.158	0.336	150704.979	-334.922	281.034	0.207		-0.363	0.716	288.858
kmeans.20	0.013	800.946	0.257	0.019	-0.977	0.127	0.121	875688.086	-387.910	84.817	0.217		-0.350	0.829	83.709
kmeans.21	0.016	753.708	0.070	0.020	-0.970	0.126	0.121	789654.288	-387.287	75.392	0.211		-0.347	0.848	74.299
kmeans.22	0.028	339.472	0.497	0.019	-0.963	0.138	0.208	205690.877	-365.010	294.485	0.200		-0.361	0.700	166.675
kmeans.23	0.023	328.968	0.526	0.013	-0.980	0.139	0.206	190876.710	-366.527	374.332	0.195		-0.368	0.721	354.944
kmeans.24	0.027	303.306	0.460	0.018	-0.925	0.135	0.209	170462.925	-364.991	188.031	0.191		-0.347	0.704	192.804
kmeans.25	0.024	603.627	0.310	0.019	-0.941	0.123	0.124	541112.909	-384.658	88.069	0.193		-0.336	0.813	87.416
kmeans.26	0.018	665.461	0.458	0.015	-0.999	0.127	0.117	564634.314	-387.573	148.069	0.190		-0.356	0.809	138.859
kmeans.27	0.031	275.203	0.733	0.016	-0.971	0.138	0.208	138373.613	-364.397	269.450	0.180		-0.363	0.689	234.329
kmeans.28	0.031	567.461	0.357	0.018	-0.987	0.125	0.122	453103.264	-383.500	301.133	0.183		-0.350	0.754	96.682
kmeans.29	0.000	186978.595	0.020	0.436	-1.000	0.113	0.012	28477214.610	-407.870	0.419	0.186		-0.336	0.779	0.242
kmeans.30	0.000	182424.263	0.263	0.380	-1.000	0.113	0.012	26865232.950	-407.878	0.354	0.183		-0.336	0.813	0.344
kmeans.31	0.000	178749.065	0.176	0.412	-1.000	0.112	0.012	25340245.500	-406.289	0.719	0.180		-0.334	0.800	0.293
kmeans.32	0.032	496.044	0.385	0.013	-0.989	0.125	0.122	351288.421	-383.171	186.848	0.171		-0.351	0.746	181.873

Tabella A.19: Dim1024

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.077	79.136	3.593	0.932	-0.253	0.313	0.873	222756.700	-157.389	3.229	0.187	0.127	-0.179	0.126	0.469
kmeans.3	0.065	85.712	3.021	0.922	-0.355	0.320	0.834	221456.468	-199.876	2.228	0.218	0.218	-0.244	0.179	0.436
kmeans.4	0.058	87.822	2.580	0.915	-0.389	0.265	0.759	148251.114	-261.354	1.665	0.227	0.252	-0.240	0.217	0.407
kmeans.5	0.045	99.265	2.701	0.919	-0.534	0.299	0.761	363736.233	-262.315	1.564	0.237	0.525	-0.333	0.301	0.369
kmeans.6	0.021	112.697	1.475	0.953	-0.735	0.303	0.714	218507.823	-297.444	1.097	0.244	0.497	-0.435	0.336	0.320
kmeans.7	0.034	91.612	0.614	0.014	-0.616	0.275	0.709	216709.765	-298.014	1531.590	0.224		-0.360	0.323	1452.091
kmeans.8	0.023	104.932	1.562	0.015	-0.745	0.273	0.673	217452.785	-321.699	1151.097	0.228		-0.417	0.386	1116.055
kmeans.9	0.023	152.178	0.683	0.935	-0.674	0.201	0.555	244469.870	-382.385	0.600	0.246	0.713	-0.330	0.534	0.230
kmeans.10	0.021	138.277	0.720	0.015	-0.774	0.242	0.614	215953.700	-353.826	946.793	0.234		-0.405	0.516	920.562
kmeans.11	0.012	213.863	0.589	0.968	-0.855	0.179	0.429	301512.883	-440.619	0.353	0.248	0.793	-0.376	0.650	0.164
kmeans.12	0.023	147.312	0.721	0.013	-0.755	0.266	0.660	239824.461	-329.032	1918.171	0.226		-0.416	0.613	959.979
kmeans.13	0.029	102.258	0.627	0.014	-0.728	0.244	0.633	132590.678	-341.964	2253.911	0.206		-0.387	0.471	1128.005
kmeans.14	0.042	162.764	0.570	0.006	-0.783	0.170	0.438	185454.734	-428.734	17923.422	0.220		-0.342	0.575	3630.723
kmeans.15	0.026	114.903	0.568	0.014	-0.764	0.211	0.555	134941.270	-381.649	1006.163	0.202		-0.374	0.520	936.527
kmeans.16	0.013	281.235	0.529	0.015	-0.899	0.147	0.278	347866.851	-497.656	432.607	0.225		-0.354	0.728	416.207
kmeans.17	0.019	260.514	0.303	0.013	-0.830	0.153	0.340	364419.690	-473.708	578.738	0.217		-0.340	0.753	542.917
kmeans.18	0.023	176.536	1.087	0.011	-0.956	0.175	0.386	177465.294	-454.558	4699.499	0.204		-0.405	0.618	1022.948
kmeans.19	0.016	237.290	0.391	0.014	-0.950	0.151	0.278	247437.119	-496.933	471.699	0.206		-0.373	0.672	448.765
kmeans.20	0.022	222.130	0.570	0.014	-0.872	0.156	0.339	267298.924	-472.996	501.657	0.201		-0.357	0.730	466.907
kmeans.21	0.000	551241.847	2.644	0.214	-0.997	0.111	0.009	203947176.600	-577.537	6.211	0.218		-0.333	0.875	2.077
kmeans.22	0.023	324.470	0.580	0.013	-0.842	0.142	0.280	527614.025	-494.092	663.323	0.199		-0.330	0.763	345.311
kmeans.23	0.022	314.540	0.161	0.014	-0.964	0.151	0.279	488765.110	-494.806	554.899	0.195		-0.378	0.729	325.170
kmeans.24	0.024	186.476	0.351	0.013	-0.960	0.150	0.279	156425.364	-494.377	582.368	0.184		-0.376	0.633	544.449
kmeans.25	0.052	177.122	0.565	0.006	-0.879	0.154	0.347	171518.383	-461.510	11229.299	0.180		-0.355	0.637	2821.646
kmeans.26	0.052	266.233	0.512	0.005	-0.883	0.129	0.216	284141.793	-507.533	7368.631	0.183		-0.327	0.704	2404.471
kmeans.27	0.051	535.705	0.203	0.007	-0.872	0.116	0.126	943568.377	-536.930	4522.144	0.186		-0.307	0.697	727.878
kmeans.28	0.032	249.677	0.400	0.010	-0.912	0.133	0.211	247769.739	-515.390	740.526	0.176		-0.341	0.668	663.635
kmeans.29	0.000	405810.088	0.428	0.553	-1.000	0.113	0.009	111118614.300	-581.711	0.385	0.186		-0.336	0.792	0.221
kmeans.30	0.062	228.151	1.199	0.005	-0.852	0.126	0.218	213794.869	-503.513	8853.515	0.170		-0.315	0.633	2981.953
kmeans.31	0.000	379805.627	0.068	0.512	-1.000	0.112	0.009	97937816.060	-580.533	0.451	0.180		-0.335	0.804	0.267
kmeans.32	0.061	476.001	0.600	0.006	-0.985	0.121	0.124	708828.803	-534.386	5433.353	0.171		-0.344	0.681	812.591

Tabella A.20: Ecoli

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
kmeans.2	0.127	235.362	0.976	0.08	-0.721	0.423	0.604	0.152	-0.138	0.308	0.347	0.427	-0.505	0.545	12.395
kmeans.3	0.163	143.607	0.961	0.07	-0.655	0.414	0.62	0.085	-0.127	0.718	0.302	0.298	-0.463	0.468	15.005
kmeans.4	0.084	203.656	1.226	0.053	-0.838	0.387	0.492	0.096	-0.15	0.565	0.329	0.337	-0.544	0.507	21.036
kmeans.5	0.136	163.068	1.218	0.064	-0.743	0.296	0.521	0.077	-0.12	1.076	0.299	0.24	-0.433	0.424	17.155
kmeans.6	0.087	161.646	1.103	0.049	-0.837	0.286	0.457	0.06	-0.13	0.74	0.292	0.249	-0.467	0.406	25.117
kmeans.7	0.1	144.375	1.317	0.049	-0.816	0.275	0.469	0.047	-0.125	0.967	0.301	0.258	-0.449	0.43	23.847
kmeans.8	0.097	137.411	1.202	0.043	-0.822	0.226	0.45	0.048	-0.115	1.171	0.28	0.203	-0.409	0.351	29.373
kmeans.9	0.098	132.797	0.975	0.048	-0.82	0.216	0.448	0.039	-0.112	0.93	0.267	0.235	-0.399	0.413	21.463
kmeans.10	0.107	112.182	1.345	0.052	-0.798	0.172	0.447	0.035	-0.099	1.04	0.231	0.19	-0.349	0.334	21.145
kmeans.11	0.076	126.004	0.869	0.061	-0.862	0.192	0.412	0.031	-0.111	0.8	0.249	0.226	-0.392	0.384	11.46
kmeans.12	0.088	116.472	1.175	0.043	-0.839	0.156	0.415	0.032	-0.098	0.813	0.24	0.204	-0.345	0.368	23.316
kmeans.13	0.077	113.562	1.231	0.054	-0.862	0.161	0.401	0.028	-0.102	1.057	0.229	0.202	-0.358	0.35	22.133
kmeans.14	0.072	113.742	1.156	0.049	-0.873	0.136	0.383	0.027	-0.095	0.846	0.225	0.196	-0.333	0.353	15.492
kmeans.15	0.066	109.347	1.29	0.049	-0.883	0.141	0.377	0.022	-0.098	0.607	0.218	0.197	-0.341	0.364	15.017
kmeans.16	0.08	102.522	1.034	0.069	-0.856	0.132	0.395	0.02	-0.092	1.16	0.212	0.213	-0.323	0.365	10.291

Tabella A.22: Aggregation Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.131	692.981	0.918	0.159	-0.710	0.424	0.539	166.907	-4.958	0.259	0.449	0.464	-0.500	0.554	4.016
upgma.3	0.065	988.520	0.657	0.135	-0.867	0.428	0.411	223.539	-5.785	0.165	0.483	0.532	-0.587	0.617	3.671
upgma.4	0.043	1141.694	0.554	0.027	-0.919	0.402	0.358	290.417	-5.891	0.156	0.453	0.558	-0.595	0.632	99.961
upgma.5	0.030	1208.002	0.531	0.030	-0.945	0.354	0.311	235.683	-5.700	0.206	0.415	0.533	-0.570	0.594	74.836
upgma.6	0.021	1306.181	0.502	0.036	-0.962	0.337	0.289	238.817	-5.663	0.158	0.384	0.566	-0.564	0.608	57.388
upgma.7	0.022	1200.172	0.435	0.036	-0.961	0.333	0.287	236.829	-5.628	0.232	0.358	0.607	-0.560	0.610	52.510
upgma.8	0.024	1104.344	0.409	0.036	-0.957	0.322	0.284	194.462	-5.525	0.317	0.336	0.585	-0.549	0.576	49.186
upgma.9	0.028	1272.239	0.434	0.039	-0.949	0.246	0.250	196.082	-4.849	0.296	0.320	0.553	-0.477	0.568	38.155
upgma.10	0.028	1288.171	0.403	0.041	-0.950	0.224	0.239	178.410	-4.652	0.303	0.305	0.534	-0.456	0.561	33.749
upgma.11	0.029	1288.038	0.413	0.048	-0.948	0.207	0.232	175.789	-4.475	0.334	0.292	0.501	-0.437	0.522	30.532
upgma.12	0.029	1346.915	0.351	0.051	-0.950	0.182	0.218	168.639	-4.213	0.308	0.281	0.491	-0.410	0.528	26.709
upgma.13	0.029	1357.029	0.356	0.051	-0.950	0.169	0.211	157.846	-4.067	0.339	0.271	0.471	-0.395	0.511	24.380
upgma.14	0.028	1364.978	0.345	0.051	-0.950	0.155	0.204	148.103	-3.912	0.362	0.261	0.450	-0.379	0.513	22.430

Tabella A.23: Wine Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.243	2.665	0.595	0.371	-0.676	0.019	0.720	8.790	-0.199	0.369	0.086		-0.101	0.476	0.966
upgma.3	0.231	4.031	0.596	0.224	-0.604	0.068	0.748	11.384	-0.334	0.544	0.121		-0.176	0.363	2.724
upgma.4	0.224	4.826	1.332	0.231	-0.568	0.125	0.762	6.620	-0.421	0.618	0.139		-0.227	0.345	2.630
upgma.5	0.146	22.274	1.278	0.255	-0.677	0.419	0.710	6.086	-0.833	0.519	0.261		-0.479	0.383	1.881
upgma.6	0.138	19.425	1.240	0.255	-0.696	0.424	0.702	4.356	-0.856	0.502	0.245		-0.492	0.373	1.821

Tabella A.24: Glass Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.133	17.548	0.295	0.493	-0.948	0.036	0.372	13.881	-0.642	0.112	0.233	0.801	-0.181	0.740	0.164
upgma.3	0.034	38.877	0.271	0.197	-0.969	0.133	0.349	14.093	-1.238	0.103	0.277	0.712	-0.356	0.700	1.453
upgma.4	0.028	29.524	0.298	0.197	-0.972	0.148	0.349	9.084	-1.288	0.119	0.254		-0.377	0.667	1.398
upgma.5	0.037	23.776	0.385	0.197	-0.960	0.162	0.356	5.973	-1.296	0.199	0.234		-0.390	0.625	1.366
upgma.6	0.055	22.486	0.378	0.226	-0.909	0.198	0.373	4.334	-1.309	0.235	0.230		-0.414	0.582	1.290
upgma.7	0.055	19.664	0.378	0.226	-0.909	0.198	0.374	3.979	-1.309	0.231	0.221		-0.414	0.583	1.266
upgma.8	0.028	53.259	0.392	0.150	-0.951	0.456	0.323	5.736	-1.650	0.171	0.251		-0.650	0.639	1.616
upgma.9	0.025	66.462	0.320	0.223	-0.959	0.464	0.316	5.668	-1.669	0.138	0.253		-0.660	0.673	1.263
upgma.10	0.020	67.320	0.321	0.225	-0.970	0.478	0.302	4.980	-1.691	0.174	0.247		-0.676	0.624	1.144
upgma.11	0.019	62.841	0.326	0.225	-0.971	0.478	0.301	4.239	-1.691	0.169	0.238		-0.676	0.624	1.109
upgma.12	0.019	58.751	0.324	0.225	-0.971	0.478	0.301	3.985	-1.691	0.165	0.229		-0.676	0.626	1.081
upgma.13	0.019	55.670	0.294	0.225	-0.971	0.478	0.301	3.497	-1.691	0.158	0.221		-0.676	0.635	1.050
upgma.14	0.019	54.152	0.296	0.225	-0.971	0.478	0.301	3.144	-1.691	0.172	0.215		-0.676	0.635	1.005

Tabella A.25: Yeast Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.114	27.834	0.419	0.322	-0.972	0.008	0.404	0.192	-0.035	0.107	0.105	0.704	-0.087	0.716	0.483
upgma.3	0.087	68.672	0.413	0.345	-0.954	0.046	0.432	0.152	-0.072	0.138	0.218	0.641	-0.207	0.680	0.450
upgma.4	0.098	57.856	0.432	0.117	-0.935	0.062	0.449	0.087	-0.079	0.209	0.196	0.564	-0.237	0.580	4.047
upgma.5	0.102	57.587	0.435	0.121	-0.928	0.095	0.472	0.057	-0.087	0.271	0.237	0.510	-0.291	0.534	3.913
upgma.6	0.110	54.497	0.437	0.121	-0.897	0.117	0.490	0.040	-0.090	0.297	0.223	0.472	-0.315	0.495	3.818
upgma.7	0.110	46.592	0.403	0.121	-0.897	0.117	0.490	0.030	-0.090	0.296	0.207	0.511	-0.315	0.490	3.802
upgma.8	0.112	41.551	0.403	0.121	-0.891	0.123	0.494	0.023	-0.091	0.308	0.198	0.485	-0.321	0.448	3.778
upgma.9	0.112	37.786	0.403	0.121	-0.891	0.123	0.494	0.019	-0.091	0.329	0.187	0.484	-0.322	0.447	3.753
upgma.10	0.112	34.303	0.403	0.121	-0.891	0.123	0.494	0.020	-0.091	0.328	0.178	0.499	-0.322	0.443	3.739
upgma.11	0.123	31.801	0.406	0.121	-0.870	0.130	0.504	0.016	-0.090	0.504	0.171	0.479	-0.324	0.373	3.719
upgma.12	0.175	69.221	0.392	0.066	-0.656	0.333	0.607	0.017	-0.092	0.428	0.189	0.460	-0.416	0.360	12.463
upgma.13	0.175	63.757	0.394	0.066	-0.656	0.333	0.607	0.014	-0.092	0.427	0.182		-0.416	0.360	12.440
upgma.14	0.171	61.099	0.392	0.066	-0.660	0.341	0.606	0.012	-0.092	0.453	0.176		-0.423	0.272	12.277
upgma.15	0.169	58.114	0.391	0.070	-0.662	0.346	0.605	0.011	-0.093	0.459	0.171		-0.428	0.255	12.170
upgma.16	0.169	54.885	0.390	0.070	-0.662	0.346	0.605	0.010	-0.093	0.457	0.167		-0.428	0.253	12.116
upgma.17	0.169	52.085	0.390	0.070	-0.663	0.349	0.605	0.009	-0.093	0.454	0.162		-0.429	0.230	12.060
upgma.18	0.146	64.743	0.382	0.079	-0.678	0.400	0.599	0.008	-0.096	0.473	0.168		-0.468	0.243	10.801
upgma.19	0.146	61.336	0.409	0.079	-0.678	0.400	0.599	0.008	-0.096	0.472	0.163		-0.468	0.246	10.784
upgma.20	0.146	58.425	0.409	0.079	-0.678	0.400	0.599	0.007	-0.096	0.470	0.159		-0.468	0.246	10.755

Tabella A.26: R15 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.244	192.535	0.945	0.22	-0.523	0.271	0.641	14.719	-1.204	0.359	0.349	0.473	-0.312	0.528	1.73
upgma.3	0.142	276.209	0.875	0.278	-0.665	0.407	0.54	16.226	-1.778	0.247	0.4	0.507	-0.465	0.507	1.188
upgma.4	0.104	287.449	0.809	0.309	-0.74	0.434	0.49	21.971	-1.947	0.228	0.384	0.58	-0.523	0.511	0.935
upgma.5	0.019	466.93	0.804	0.387	-0.953	0.439	0.347	21.066	-2.291	0.154	0.389	0.585	-0.639	0.572	0.674
upgma.6	0.014	494.961	0.397	0.387	-0.965	0.436	0.338	20.388	-2.302	0.134	0.367	0.695	-0.643	0.625	0.54
upgma.7	0.01	575.633	0.398	0.387	-0.977	0.432	0.328	23.31	-2.313	0.104	0.349	0.776	-0.646	0.679	0.409
upgma.8	0.005	760.074	0.354	0.494	-0.989	0.429	0.317	28.817	-2.325	0.072	0.335	0.843	-0.649	0.739	0.279
upgma.9	0.04	975.165	0.32	0.06	-0.932	0.318	0.284	32.877	-1.962	0.237	0.321	0.792	-0.534	0.681	13.954
upgma.10	0.022	1556.153	0.28	0.068	-0.956	0.23	0.207	41.611	-1.755	0.12	0.31	0.765	-0.464	0.71	8.01
upgma.11	0.015	1957.968	0.279	0.092	-0.968	0.192	0.171	46.302	-1.64	0.13	0.297	0.745	-0.428	0.713	5.787
upgma.12	0.012	2277.156	0.272	0.094	-0.975	0.165	0.145	49.526	-1.542	0.163	0.285	0.733	-0.398	0.737	4.545
upgma.13	0.01	2711.083	0.194	0.1	-0.981	0.151	0.129	55.649	-1.489	0.126	0.275	0.739	-0.383	0.756	3.512
upgma.14	0.007	3458.172	0.172	0.106	-0.988	0.136	0.111	65.582	-1.434	0.091	0.266	0.749	-0.366	0.791	2.55
upgma.15	0.001	4860.667	0.166	0.186	-0.998	0.122	0.088	78.819	-1.377	0.065	0.257	0.752	-0.348	0.821	1.689
upgma.16	0.001	4573.414	0.166	0.186	-0.999	0.121	0.088	70.015	-1.375	0.167	0.249		-0.348	0.804	1.673
upgma.17	0.001	4324.069	0.166	0.195	-0.999	0.121	0.088	62.56	-1.373	0.165	0.242		-0.347	0.793	1.656
upgma.18	0.001	4176.771	0.161	0.195	-0.999	0.12	0.087	57.281	-1.367	0.166	0.235		-0.346	0.786	1.611
upgma.19	0.001	3973.621	0.162	0.195	-0.998	0.119	0.087	56.522	-1.365	0.181	0.228		-0.345	0.767	1.597
upgma.20	0.001	3820.384	0.16	0.195	-0.998	0.119	0.086	51.773	-1.361	0.192	0.223		-0.344	0.75	1.571
upgma.21	0.001	3673.483	0.159	0.195	-0.998	0.118	0.086	47.566	-1.357	0.229	0.217		-0.343	0.732	1.549
upgma.22	0.002	3602.955	0.216	0.13	-0.998	0.116	0.085	44.466	-1.346	0.271	0.212		-0.34	0.709	3.388
upgma.23	0.002	3458.359	0.211	0.13	-0.998	0.116	0.084	41.017	-1.344	0.269	0.208		-0.34	0.698	3.364
upgma.24	0.002	3345.017	0.208	0.13	-0.998	0.115	0.084	38.104	-1.339	0.266	0.203		-0.338	0.687	3.321
upgma.25	0.002	3241.617	0.206	0.13	-0.998	0.114	0.084	35.506	-1.335	0.262	0.199		-0.337	0.665	3.279
upgma.26	0.002	3178.521	0.204	0.13	-0.998	0.113	0.083	35.212	-1.327	0.261	0.195		-0.335	0.642	3.205
upgma.27	0.002	3071.25	0.203	0.13	-0.998	0.112	0.083	32.885	-1.324	0.259	0.192		-0.335	0.636	3.184
upgma.28	0.002	2989.767	0.2	0.13	-0.997	0.112	0.083	31.017	-1.321	0.256	0.188		-0.334	0.632	3.145
upgma.29	0.002	2897.42	0.199	0.13	-0.997	0.111	0.082	30.574	-1.319	0.252	0.185		-0.333	0.629	3.124
upgma.30	0.002	2860.782	0.197	0.13	-0.997	0.11	0.082	29.278	-1.31	0.248	0.182		-0.331	0.611	3.05

Tabella A.27: Iris Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.023	501.92	0.384	0.339	-0.958	0.484	0.324	20.400	-1.365	0.066	0.560	0.722	-0.674	0.774	0.384
upgma.3	0.033	555.66	0.616	0.138	-0.916	0.434	0.274	25.378	-1.175	0.162	0.499	0.559	-0.616	0.666	3.787
upgma.4	0.026	433.51	0.671	0.154	-0.926	0.423	0.261	16.215	-1.159	0.159	0.439	0.547	-0.613	0.598	3.272
upgma.5	0.024	397.42	0.593	0.154	-0.926	0.404	0.253	15.770	-1.120	0.257	0.398	0.494	-0.600	0.563	2.710
upgma.6	0.024	325.50	0.525	0.167	-0.926	0.400	0.252	12.911	-1.113	0.250	0.366		-0.597	0.503	2.635

Tabella A.28: Flame Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.196	151.64	1.133	0.049	-0.588	0.397	0.613	14.905	-1.427	0.391	0.436	0.374	-0.416	0.495	28.544
upgma.3	0.137	172.45	0.691	0.056	-0.711	0.402	0.527	15.086	-1.661	0.228	0.443	0.447	-0.488	0.523	19.033
upgma.4	0.071	245.549	0.66	0.062	-0.856	0.374	0.422	26.356	-1.844	0.153	0.435	0.468	-0.544	0.566	12.339
upgma.5	0.067	215.277	0.958	0.081	-0.866	0.349	0.404	18.754	-1.8	0.327	0.396	0.421	-0.53	0.519	10.903
upgma.6	0.068	175.641	0.684	0.081	-0.866	0.348	0.404	21.647	-1.798	0.446	0.363	0.454	-0.529	0.5	10.699

Tabella A.29: Pathbased Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.128	316.918	0.801	0.049	-0.74	0.435	0.505	133.865	-4.168	0.234	0.446	0.496	-0.523	0.623	24.825
upgma.3	0.052	353.957	0.621	0.054	-0.89	0.458	0.404	113.598	-4.933	0.127	0.467	0.528	-0.62	0.657	18.117
upgma.4	0.046	286.421	0.468	0.062	-0.903	0.449	0.387	182.427	-4.933	0.22	0.422	0.539	-0.621	0.606	15.707
upgma.5	0.046	241.516	0.441	0.062	-0.903	0.447	0.385	152.647	-4.923	0.233	0.382	0.539	-0.619	0.562	14.34
upgma.6	0.046	232.041	0.421	0.062	-0.906	0.43	0.371	122.303	-4.835	0.248	0.359	0.515	-0.609	0.541	12.393

Tabella A.30: Dim032 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.174	184.536	2.227	0.564	-0.436	0.358	0.825	18151.25	-40.314	1.298	0.278	0.205	-0.307	0.247	0.805
upgma.3	0.132	167.806	1.586	0.597	-0.583	0.382	0.772	15651.348	-51.613	1.04	0.287	0.304	-0.405	0.283	0.715
upgma.4	0.077	188.241	1.672	0.635	-0.766	0.362	0.689	12025.818	-64.766	0.697	0.297	0.341	-0.491	0.336	0.612
upgma.5	0.06	192.67	0.692	0.656	-0.83	0.342	0.649	17415.243	-69.487	0.612	0.294	0.461	-0.508	0.393	0.541
upgma.6	0.052	202.602	0.692	0.659	-0.859	0.333	0.631	14694.083	-71.503	0.492	0.288	0.571	-0.514	0.455	0.476
upgma.7	0.049	213.909	0.692	0.659	-0.867	0.326	0.621	17901.933	-72.432	0.434	0.281	0.687	-0.513	0.523	0.42
upgma.8	0.03	256.084	0.692	0.709	-0.932	0.284	0.549	16814.418	-78.665	0.338	0.282	0.686	-0.505	0.561	0.344
upgma.9	0.024	298.274	0.602	0.779	-0.965	0.225	0.443	15945.307	-85.119	0.353	0.278	0.679	-0.462	0.582	0.284
upgma.10	0.017	351.179	0.602	0.779	-0.976	0.203	0.388	17115.829	-88.278	0.288	0.273	0.727	-0.442	0.638	0.231
upgma.11	0.007	426.362	1.019	0.824	-0.992	0.18	0.318	18951.894	-92.131	0.193	0.269	0.76	-0.422	0.69	0.182
upgma.12	0.004	528.128	0.464	0.867	-0.996	0.168	0.277	24336.904	-94.019	0.149	0.265	0.808	-0.409	0.747	0.141
upgma.13	0.002	694.258	0.487	0.881	-0.998	0.155	0.232	34941.088	-95.847	0.1	0.261	0.853	-0.394	0.808	0.103
upgma.14	0.001	1033.514	0.491	0.901	-0.999	0.142	0.179	67900.516	-97.857	0.065	0.257	0.889	-0.377	0.863	0.066
upgma.15	0	2063.91	0.381	0.942	-1	0.129	0.116	158833.618	-100.086	0.029	0.253	0.922	-0.359	0.918	0.032
upgma.16	0	27667.061	0.063	4.035	-1	0.116	0.043	818251.881	-102.208	0.002	0.25	0.946	-0.34	0.965	0.003
upgma.17	0	26258.766	0.063	0.899	-1	0.115	0.043	731385.152	-102.121	0.072	0.242		-0.34	0.952	0.079
upgma.18	0	24864.536	0.063	0.631	-1	0.115	0.042	656771.296	-102.027	0.136	0.235		-0.339	0.936	0.16
upgma.19	0	23608.164	0.062	0.594	-1	0.115	0.042	593051.846	-101.932	0.152	0.229		-0.339	0.916	0.18
upgma.20	0	22489.866	0.062	0.594	-1	0.115	0.042	538546.039	-101.837	0.151	0.223		-0.339	0.9	0.178
upgma.21	0	21485.279	0.061	0.594	-1	0.115	0.042	491477.37	-101.744	0.15	0.218		-0.338	0.895	0.177
upgma.22	0	20582.937	0.061	0.59	-1	0.114	0.042	450718.837	-101.648	0.149	0.213		-0.338	0.88	0.178
upgma.23	0	19765.423	0.062	0.613	-1	0.114	0.042	415023.203	-101.553	0.148	0.208		-0.338	0.867	0.177
upgma.24	0	19007.728	0.061	0.629	-1	0.114	0.042	383579.962	-101.458	0.151	0.204		-0.337	0.867	0.176
upgma.25	0	18305.039	0.061	0.612	-1	0.114	0.042	355591.86	-101.364	0.163	0.2		-0.337	0.866	0.185
upgma.26	0	17670.454	0.061	0.612	-1	0.113	0.042	330772.386	-101.267	0.162	0.196		-0.337	0.852	0.184
upgma.27	0	17076.95	0.06	0.636	-1	0.113	0.042	308556.001	-101.174	0.161	0.192		-0.337	0.852	0.187
upgma.28	0	16516.173	0.06	0.621	-1	0.113	0.042	288515.585	-101.076	0.179	0.189		-0.336	0.831	0.195
upgma.29	0	15995.939	0.06	0.608	-1	0.113	0.041	270473.24	-100.98	0.179	0.185		-0.336	0.831	0.203
upgma.30	0	15512.48	0.06	0.608	-1	0.113	0.041	254275.667	-100.886	0.178	0.182		-0.336	0.829	0.202
upgma.31	0	15063.43	0.059	0.608	-1	0.112	0.041	239516.574	-100.792	0.177	0.179		-0.335	0.829	0.2
upgma.32	0	14644.78	0.059	0.608	-1	0.112	0.041	226098.006	-100.694	0.176	0.177		-0.335	0.811	0.199

Tabella A.31: Jain toy Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.149	294.007	0.778	0.045	-0.67	0.368	0.529	165.08	-4.718	0.207	0.49	0.487	-0.445	0.502	52.547
upgma.3	0.075	522.052	0.62	0.041	-0.836	0.423	0.389	306.916	-5.484	0.173	0.481	0.473	-0.568	0.561	35.825
upgma.4	0.071	450.13	0.516	0.051	-0.847	0.415	0.376	315.522	-5.477	0.211	0.428	0.509	-0.568	0.567	29.384
upgma.5	0.061	541.897	0.464	0.048	-0.871	0.345	0.322	310.853	-5.093	0.243	0.408	0.482	-0.529	0.579	27.525
upgma.6	0.038	715.266	0.354	0.034	-0.92	0.29	0.259	381.517	-4.862	0.156	0.387	0.495	-0.506	0.581	41.797

Tabella A.32: Transfusion Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.043	208.232	0.332	0.178	-0.993	0.041	0.236	23.129	-1.013	0.038	0.33	0.753	-0.202	0.816	0.818
upgma.3	0.008	132.152	0.179	0.246	-0.997	0.051	0.238	21.203	-1.103	0.045	0.295	0.812	-0.226	0.79	0.773
upgma.4	0.079	98.662	0.184	0.193	-0.921	0.073	0.289	12.278	-1.019	0.209	0.267	0.744	-0.254	0.588	1.545
upgma.5	0.075	178.788	0.17	0.076	-0.842	0.277	0.407	10.329	-1.037	0.182	0.313	0.66	-0.461	0.514	15.134
upgma.6	0.075	145.163	0.161	0.076	-0.842	0.277	0.407	8.277	-1.037	0.314	0.287	0.584	-0.461	0.511	15.014

Tabella A.33: Dim064 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.163	120.131	2.653	0.68	-0.401	0.343	0.862	23545.741	-42.231	1.828	0.222	0.178	-0.281	0.195	0.598
upgma.3	0.099	136.77	2.264	0.758	-0.617	0.36	0.77	16324.108	-66.83	1.276	0.259	0.221	-0.412	0.237	0.527
upgma.4	0.092	132.604	0.743	0.758	-0.642	0.347	0.75	25983.131	-70.616	1.164	0.261	0.412	-0.418	0.291	0.481
upgma.5	0.079	140.627	0.711	0.808	-0.701	0.325	0.712	21443.982	-77.263	0.999	0.262	0.454	-0.433	0.335	0.431
upgma.6	0.059	153.821	0.715	0.808	-0.776	0.27	0.623	17890.946	-89.837	0.715	0.263	0.463	-0.428	0.38	0.381
upgma.7	0.053	167.134	0.717	0.808	-0.794	0.254	0.593	20665.704	-93.427	0.606	0.262	0.557	-0.423	0.439	0.336
upgma.8	0.036	189.697	0.618	0.826	-0.868	0.225	0.515	18335.647	-102.397	0.458	0.262	0.581	-0.425	0.482	0.29
upgma.9	0.025	216.223	0.618	0.826	-0.901	0.207	0.462	21932.417	-108.038	0.369	0.262	0.644	-0.42	0.537	0.247
upgma.10	0.012	255.936	0.476	0.88	-0.965	0.19	0.392	22863.083	-114.874	0.273	0.261	0.693	-0.424	0.6	0.204
upgma.11	0.005	309.17	0.45	0.911	-0.986	0.18	0.35	27387.061	-118.772	0.22	0.26	0.76	-0.42	0.668	0.165
upgma.12	0.002	383.697	0.454	0.94	-0.995	0.168	0.305	35398.203	-122.452	0.16	0.258	0.811	-0.408	0.736	0.129
upgma.13	0.001	503.251	0.391	0.949	-0.997	0.155	0.254	50968.002	-126.207	0.119	0.256	0.858	-0.393	0.799	0.096
upgma.14	0	748.91	0.38	0.968	-0.999	0.142	0.193	90766.294	-130.525	0.075	0.253	0.9	-0.377	0.86	0.063
upgma.15	0	1488.164	0.387	0.986	-1	0.129	0.119	238833.094	-135.418	0.032	0.252	0.936	-0.359	0.923	0.031
upgma.16	0	54451.234	0.038	5.82	-1	0.116	0.03	2664252.09	-140.874	0.001	0.25	0.966	-0.34	0.978	0.001
upgma.17	0	52092.055	0.038	1.013	-1	0.115	0.03	2389220.81	-140.753	0.045	0.242		-0.34	0.971	0.046
upgma.18	0	49643.355	0.038	0.869	-1	0.115	0.029	2151346.75	-140.625	0.071	0.236		-0.339	0.96	0.074
upgma.19	0	47417.285	0.038	0.811	-1	0.115	0.029	1948154.42	-140.495	0.078	0.229		-0.339	0.948	0.084
upgma.20	0	45381.704	0.038	0.797	-1	0.115	0.029	1773525.44	-140.366	0.086	0.223		-0.339	0.946	0.093
upgma.21	0	43391.048	0.038	0.688	-1	0.115	0.029	1619520.69	-140.233	0.129	0.218		-0.338	0.923	0.14
upgma.22	0	41591.072	0.038	0.688	-1	0.114	0.029	1485866.78	-140.101	0.129	0.213		-0.338	0.923	0.139
upgma.23	0	39927.426	0.038	0.662	-1	0.114	0.029	1368224.65	-139.97	0.143	0.208		-0.338	0.923	0.149
upgma.24	0	38407.933	0.038	0.69	-1	0.114	0.029	1264694.35	-139.841	0.144	0.204		-0.338	0.923	0.148
upgma.25	0	37013.339	0.038	0.676	-1	0.114	0.029	1172904.98	-139.713	0.146	0.2		-0.337	0.924	0.153
upgma.26	0	35718.865	0.038	0.661	-1	0.114	0.029	1091161.45	-139.588	0.154	0.196		-0.337	0.924	0.159
upgma.27	0	34541.525	0.038	0.661	-1	0.113	0.029	1018459.42	-139.451	0.153	0.192		-0.337	0.906	0.158
upgma.28	0	33436.648	0.038	0.639	-1	0.113	0.029	952890.21	-139.327	0.154	0.189		-0.336	0.906	0.168
upgma.29	0	32441.837	0.038	0.655	-1	0.113	0.029	894338.199	-139.192	0.153	0.186		-0.336	0.902	0.167
upgma.30	0	31489.898	0.038	0.675	-1	0.113	0.029	840794.278	-139.055	0.152	0.182		-0.336	0.882	0.167
upgma.31	0	30603.234	0.037	0.675	-1	0.112	0.028	792475.784	-138.917	0.151	0.179		-0.335	0.865	0.166
upgma.32	0	29768.254	0.037	0.675	-1	0.112	0.028	748393.616	-138.779	0.156	0.177		-0.335	0.845	0.165

Tabella A.34: Dim128 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.05	87.394	0.869	0.85	-0.648	0.171	0.886	134260.468	-32.969	0.685	0.199	0.546	-0.295	0.39	0.489
upgma.3	0.053	95.273	0.856	0.852	-0.694	0.297	0.872	122486.123	-48.047	0.626	0.228	0.686	-0.411	0.422	0.448
upgma.4	0.062	106.714	0.759	0.848	-0.692	0.411	0.849	76916.426	-65.78	0.999	0.243	0.64	-0.482	0.362	0.408
upgma.5	0.063	113.743	0.768	0.865	-0.681	0.42	0.832	60692.712	-73.729	0.904	0.247	0.705	-0.481	0.409	0.37
upgma.6	0.054	130.217	0.752	0.894	-0.7	0.365	0.761	47791.689	-96.346	1.022	0.253	0.651	-0.459	0.398	0.333
upgma.7	0.031	150.288	0.683	0.886	-0.826	0.311	0.674	40466.315	-116.939	0.808	0.256	0.639	-0.482	0.44	0.312
upgma.8	0.029	165.86	0.625	0.886	-0.836	0.305	0.663	42157.673	-119.109	0.712	0.256	0.74	-0.482	0.513	0.274
upgma.9	0.017	190.857	0.602	0.9	-0.907	0.27	0.595	42372.475	-132.022	0.564	0.256	0.77	-0.483	0.57	0.235
upgma.10	0.011	226.865	0.657	0.905	-0.944	0.255	0.559	45711.799	-138.474	0.373	0.257	0.809	-0.484	0.639	0.195
upgma.11	0.007	279.235	0.596	0.927	-0.987	0.192	0.399	46366.146	-161.137	0.304	0.257	0.792	-0.434	0.669	0.162
upgma.12	0.001	351.357	0.494	0.969	-0.997	0.168	0.311	55066.831	-172.236	0.196	0.255	0.82	-0.409	0.727	0.126
upgma.13	0.001	464.651	0.479	0.969	-0.998	0.155	0.257	81325.673	-178.288	0.13	0.254	0.869	-0.394	0.796	0.093
upgma.14	0	698.166	0.423	0.998	-1	0.143	0.193	148720.191	-185.202	0.085	0.253	0.91	-0.378	0.861	0.061
upgma.15	0	1358.692	0.432	0.982	-0.999	0.129	0.118	405285.107	-192.576	0.033	0.251	0.945	-0.359	0.928	0.032
upgma.16	0	86413.57	0.049	7.48	-1	0.116	0.023	7119600.46	-201.464	0	0.25	0.975	-0.34	0.983	0
upgma.17	0	82851.132	0.049	0.806	-1	0.115	0.023	6418142.2	-201.285	0.041	0.242		-0.34	0.975	0.042
upgma.18	0	79598.228	0.048	0.922	-1	0.115	0.023	5792281.28	-201.106	0.044	0.236		-0.339	0.962	0.046
upgma.19	0	76240.267	0.047	0.838	-1	0.115	0.023	5250179.58	-200.922	0.063	0.229		-0.339	0.96	0.066
upgma.20	0	73361.963	0.047	0.838	-1	0.115	0.023	4793951.51	-200.736	0.062	0.224		-0.339	0.954	0.065
upgma.21	0	70510.425	0.052	0.762	-1	0.115	0.023	4387175.65	-200.546	0.076	0.218		-0.338	0.937	0.078
upgma.22	0	67947.381	0.051	0.761	-1	0.114	0.023	4034518.83	-200.355	0.075	0.213		-0.338	0.927	0.077
upgma.23	0	65584.547	0.05	0.74	-1	0.114	0.022	3724461.63	-200.163	0.079	0.208		-0.338	0.912	0.081
upgma.24	0	63386.498	0.048	0.713	-1	0.114	0.022	3449659.87	-199.978	0.084	0.204		-0.337	0.91	0.086
upgma.25	0	61403.641	0.047	0.713	-1	0.114	0.022	3206759.2	-199.785	0.083	0.2		-0.337	0.895	0.085
upgma.26	0	59525.984	0.046	0.683	-1	0.113	0.022	2991627.96	-199.594	0.089	0.196		-0.337	0.891	0.092
upgma.27	0	57741.879	0.045	0.711	-1	0.113	0.022	2795882.6	-199.409	0.098	0.192		-0.337	0.89	0.099
upgma.28	0	56095.982	0.043	0.737	-1	0.113	0.022	2619748.62	-199.226	0.097	0.189		-0.336	0.891	0.101
upgma.29	0	54537.798	0.042	0.735	-1	0.113	0.022	2460644.35	-199.046	0.104	0.186		-0.336	0.891	0.109
upgma.30	0	53102.765	0.041	0.786	-1	0.113	0.022	2316678.71	-198.853	0.103	0.183		-0.336	0.89	0.108
upgma.31	0	51689.941	0.04	0.738	-1	0.112	0.022	2184921.42	-198.673	0.12	0.18		-0.335	0.889	0.121
upgma.32	0	50412.309	0.04	0.738	-1	0.112	0.022	2066617.96	-198.478	0.119	0.177		-0.335	0.887	0.12

Tabella A.35: Dim256 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.069	88.546	2.004	0.873	-0.502	0.257	0.896	142699.931	-54.295	1.262	0.201	0.299	-0.294	0.257	0.468
upgma.3	0.08	95.113	1.739	0.893	-0.5	0.362	0.872	117253.579	-78.841	1.168	0.229	0.37	-0.348	0.269	0.437
upgma.4	0.08	99.379	0.812	0.893	-0.518	0.379	0.854	105416.647	-91.064	1.049	0.238	0.522	-0.366	0.32	0.401
upgma.5	0.055	112.498	0.778	0.875	-0.634	0.335	0.761	74891.892	-134.567	1.377	0.247	0.478	-0.406	0.303	0.386
upgma.6	0.043	122	0.726	0.875	-0.696	0.311	0.719	82942.711	-149.778	1.166	0.249	0.567	-0.421	0.362	0.348
upgma.7	0.041	133.212	0.737	0.875	-0.707	0.306	0.71	80227.271	-152.664	1.044	0.25	0.696	-0.423	0.437	0.311
upgma.8	0.031	152.539	0.679	0.88	-0.802	0.256	0.615	70635.113	-179.944	0.746	0.253	0.682	-0.427	0.483	0.279
upgma.9	0.021	175.564	0.658	0.88	-0.861	0.224	0.535	64848.649	-199.781	0.483	0.254	0.689	-0.422	0.532	0.24
upgma.10	0.019	202.714	0.542	0.88	-0.866	0.214	0.51	73715.583	-205.48	0.411	0.254	0.766	-0.415	0.604	0.205
upgma.11	0.009	249.888	0.446	0.951	-0.98	0.179	0.37	74379.01	-234.652	0.265	0.254	0.766	-0.417	0.652	0.165
upgma.12	0.007	310.355	0.454	0.957	-0.985	0.167	0.322	106243.821	-243.561	0.208	0.253	0.823	-0.404	0.723	0.131
upgma.13	0.005	412.614	0.459	0.96	-0.99	0.155	0.265	158107.952	-253.522	0.154	0.253	0.873	-0.391	0.789	0.097
upgma.14	0.002	621.767	0.458	0.97	-0.995	0.142	0.197	294901.057	-264.849	0.095	0.252	0.915	-0.376	0.86	0.064
upgma.15	0	1271.989	0.277	1.006	-1	0.129	0.115	898615.401	-277.806	0.036	0.251	0.952	-0.359	0.93	0.031
upgma.16	0	203865.162	0.022	14.85	-1	0.116	0.016	30072605.6	-291.899	0	0.25	0.983	-0.34	0.989	0
upgma.17	0	193303.567	0.022	0.873	-1	0.115	0.016	26864475.7	-291.614	0.077	0.242		-0.34	0.973	0.079
upgma.18	0	183506.643	0.022	0.766	-1	0.115	0.016	24139720	-291.33	0.1	0.236		-0.339	0.972	0.101
upgma.19	0	174583.388	0.022	0.748	-1	0.115	0.016	21813268.1	-291.04	0.115	0.229		-0.339	0.953	0.117
upgma.20	0	166607.068	0.022	0.789	-1	0.115	0.016	19820645.5	-290.754	0.115	0.224		-0.339	0.953	0.117
upgma.21	0	159465.824	0.022	0.789	-1	0.115	0.016	18101638.9	-290.463	0.114	0.218		-0.338	0.939	0.116
upgma.22	0	152749.018	0.022	0.72	-1	0.114	0.016	16594301.9	-290.179	0.142	0.213		-0.338	0.938	0.143
upgma.23	0	146670.131	0.022	0.716	-1	0.114	0.016	15276630.3	-289.886	0.141	0.208		-0.338	0.918	0.143
upgma.24	0	141063.956	0.021	0.707	-1	0.114	0.016	14113795.9	-289.606	0.148	0.204		-0.338	0.918	0.15
upgma.25	0	135964.651	0.021	0.707	-1	0.114	0.016	13089481.9	-289.311	0.147	0.2		-0.337	0.901	0.149
upgma.26	0	131230.824	0.021	0.698	-1	0.113	0.016	12173818.6	-289.026	0.15	0.196		-0.337	0.898	0.152
upgma.27	0	126834.427	0.021	0.681	-1	0.113	0.016	11354667.3	-288.735	0.156	0.192		-0.337	0.897	0.159
upgma.28	0	122763.456	0.021	0.681	-1	0.113	0.016	10619423.2	-288.458	0.156	0.189		-0.336	0.897	0.158
upgma.29	0	118977.171	0.021	0.687	-1	0.113	0.016	9956365.83	-288.185	0.158	0.186		-0.336	0.898	0.163
upgma.30	0	115469.604	0.021	0.718	-1	0.113	0.015	9358049.64	-287.889	0.157	0.183		-0.336	0.877	0.162
upgma.31	0	112187.426	0.021	0.716	-1	0.112	0.015	8815129.95	-287.621	0.158	0.18		-0.335	0.878	0.161
upgma.32	0	109156.115	0.021	0.716	-1	0.112	0.015	8323960.34	-287.328	0.157	0.177		-0.335	0.878	0.16

Tabella A.36: Breast Tissue Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0	359.31	0.036	3.35	-1	0.037	0.047	1.3528E+10	-21880.943	0.003	0.318		-0.192	0.962	0.004
upgma.3	0.002	827.826	0.031	0.521	-0.997	0.244	0.097	1.5818E+10	-15385.091	0.022	0.379		-0.492	0.862	0.262
upgma.4	0.004	599.807	0.022	0.261	-0.992	0.246	0.098		-15284.389	0.176	0.33		-0.493	0.822	0.958
upgma.5	0.009	1798.282	0.016	0.074	-0.958	0.484	0.102	2.4434E+10	-9781.049	0.056	0.335		-0.674	0.769	8.494
upgma.6	0.008	1721.725	0.014	0.104	-0.959	0.486	0.1	1.9792E+10	-9654.564	0.119	0.309		-0.675	0.717	7.044
upgma.7	0.008	1476.8	0.013	0.104	-0.959	0.486	0.1	1.5238E+10	-9649.528	0.16	0.287		-0.675	0.716	6.778
upgma.8	0.01	1506.603	0.01	0.122	-0.948	0.486	0.1	1.4317E+10	-9406.087	0.218	0.27		-0.669	0.633	5.648
upgma.9	0.004	3568.308	0.01	0.121	-0.972	0.422	0.06	3.1754E+10	-6729.314	0.088	0.268		-0.635	0.7	4.01
upgma.10	0.004	3420.401	0.008	0.122	-0.972	0.421	0.06	2.7773E+10	-6724.042	0.081	0.257		-0.636	0.701	3.681
upgma.11	0.002	4175.246	0.008	0.122	-0.984	0.405	0.052	3.1267E+10	-6441.905	0.147	0.247		-0.629	0.694	2.688
upgma.12	0.002	4630.946	0.007	0.185	-0.988	0.401	0.051	3.0742E+10	-6382.9	0.119	0.241		-0.628	0.7	2.181

Tabella A.37: Dim512 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.028	81.163	0.898	0.932	-0.643	0.17	0.904	514300.548	-55.056	0.738	0.191	0.539	-0.293	0.377	0.446
upgma.3	0.052	88.943	0.882	0.927	-0.579	0.344	0.887	345072.763	-93.326	1.233	0.222	0.526	-0.382	0.296	0.423
upgma.4	0.052	100.228	0.839	0.889	-0.605	0.373	0.812	213637.791	-158.916	1.873	0.238	0.456	-0.412	0.251	0.417
upgma.5	0.034	111.504	0.752	0.907	-0.732	0.317	0.724	151423.956	-207.323	1.248	0.246	0.443	-0.443	0.3	0.376
upgma.6	0.033	119.725	0.753	0.907	-0.745	0.312	0.715	179120.966	-211.462	1.13	0.248	0.615	-0.446	0.377	0.34
upgma.7	0.026	131.98	0.728	0.917	-0.809	0.292	0.673	162800.258	-229.976	0.915	0.25	0.674	-0.459	0.439	0.304
upgma.8	0.02	148.121	0.728	0.922	-0.851	0.272	0.633	154088.022	-246.204	0.79	0.251	0.724	-0.461	0.497	0.268
upgma.9	0.014	168.861	0.728	0.932	-0.904	0.25	0.582	153618.111	-265.276	0.616	0.252	0.762	-0.463	0.555	0.232
upgma.10	0.009	197.338	0.637	0.949	-0.945	0.211	0.485	147182.301	-296.964	0.449	0.252	0.76	-0.441	0.6	0.196
upgma.11	0.006	237.179	0.629	0.949	-0.966	0.19	0.416	163013.136	-317.48	0.321	0.252	0.795	-0.425	0.658	0.162
upgma.12	0.003	299.115	0.555	0.973	-0.993	0.168	0.324	193258.562	-342.707	0.224	0.252	0.825	-0.408	0.719	0.127
upgma.13	0.002	398.902	0.511	0.98	-0.997	0.155	0.266	289625.329	-357.396	0.155	0.252	0.875	-0.393	0.79	0.094
upgma.14	0	601.426	0.514	0.997	-1	0.143	0.197	545722.147	-373.96	0.101	0.252	0.918	-0.378	0.859	0.062
upgma.15	0	1201.176	0.48	1.007	-1	0.129	0.115	1680087.95	-392.397	0.037	0.251	0.954	-0.359	0.931	0.031
upgma.16	0	330337.86	0.017	20.008	-1	0.116	0.013	87532562.7	-413.753	0	0.25	0.987	-0.34	0.991	0
upgma.17	0	312888.39	0.017	0.758	-1	0.115	0.013	78146757.6	-413.343	0.085	0.243		-0.34	0.972	0.085
upgma.18	0	297818.927	0.017	0.758	-1	0.115	0.013	70280382.2	-412.932	0.084	0.236		-0.339	0.959	0.084
upgma.19	0	283887.604	0.018	0.748	-1	0.115	0.013	63550627	-412.518	0.093	0.229		-0.339	0.944	0.095
upgma.20	0	271146.218	0.017	0.769	-1	0.115	0.013	57758691.7	-412.11	0.105	0.224		-0.339	0.943	0.105
upgma.21	0	259815.383	0.018	0.769	-1	0.115	0.012	52773568	-411.701	0.104	0.218		-0.338	0.943	0.104
upgma.22	0	249269.162	0.017	0.739	-1	0.114	0.012	48406772.4	-411.298	0.115	0.213		-0.338	0.943	0.116
upgma.23	0	239579.648	0.017	0.762	-1	0.114	0.012	44576448	-410.9	0.122	0.208		-0.338	0.943	0.122
upgma.24	0	230472.628	0.017	0.7	-1	0.114	0.012	41183414.5	-410.507	0.143	0.204		-0.338	0.942	0.143
upgma.25	0	222109.93	0.017	0.741	-1	0.114	0.012	38179450.5	-410.12	0.145	0.2		-0.337	0.943	0.145
upgma.26	0	214374.969	0.017	0.734	-1	0.114	0.012	35504895.4	-409.739	0.15	0.196		-0.337	0.943	0.149
upgma.27	0	207227.624	0.017	0.727	-1	0.113	0.012	33114998.6	-409.364	0.151	0.192		-0.337	0.944	0.151
upgma.28	0	200646.764	0.017	0.727	-1	0.113	0.012	30971222	-408.942	0.151	0.189		-0.336	0.923	0.15
upgma.29	0	194578.226	0.017	0.727	-1	0.113	0.012	29044887.5	-408.519	0.15	0.186		-0.336	0.904	0.149
upgma.30	0	188864.756	0.017	0.719	-1	0.113	0.012	27297112.2	-408.15	0.152	0.183		-0.336	0.905	0.152
upgma.31	0	183604.922	0.017	0.727	-1	0.113	0.012	25716298.5	-407.726	0.151	0.18		-0.335	0.886	0.151
upgma.32	0	178658.071	0.017	0.727	-1	0.112	0.012	24305283.3	-407.309	0.15	0.177		-0.335	0.881	0.15

Tabella A.38: Dim1024 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.021	79.397	0.905	0.945	-0.709	0.177	0.91	1015732.27	-73.376	0.754	0.19	0.538	-0.323	0.373	0.445
upgma.3	0.034	88.283	0.893	0.949	-0.617	0.352	0.893	706268.104	-125.003	1.226	0.223	0.53	-0.407	0.298	0.415
upgma.4	0.033	96.408	0.846	0.942	-0.641	0.381	0.818	433834.274	-218.159	2.039	0.235	0.458	-0.437	0.242	0.385
upgma.5	0.029	104.191	0.846	0.944	-0.668	0.363	0.793	306768.336	-239.934	1.337	0.241	0.493	-0.441	0.301	0.351
upgma.6	0.023	113.278	0.643	0.956	-0.72	0.335	0.753	311317.8	-270.958	1.157	0.244	0.577	-0.449	0.358	0.318
upgma.7	0.009	127.581	0.589	0.969	-0.897	0.269	0.625	255635.056	-350.617	0.849	0.247	0.558	-0.478	0.399	0.288
upgma.8	0.006	142.464	0.606	0.969	-0.934	0.254	0.589	250397.089	-369.895	0.751	0.248	0.633	-0.479	0.457	0.255
upgma.9	0.003	162.543	0.589	0.984	-0.979	0.227	0.519	245239.925	-403.862	0.567	0.249	0.682	-0.469	0.522	0.221
upgma.10	0.002	188.025	0.599	0.985	-0.986	0.216	0.491	279156.769	-416.795	0.485	0.249	0.761	-0.46	0.595	0.189
upgma.11	0.002	224.097	0.595	0.988	-0.992	0.193	0.422	306954.578	-446.302	0.337	0.25	0.797	-0.436	0.656	0.157
upgma.12	0.001	278.787	0.602	0.991	-0.996	0.181	0.379	411809.677	-463.532	0.269	0.25	0.854	-0.424	0.729	0.125
upgma.13	0	370.625	0.608	0.994	-0.998	0.168	0.329	645479.778	-482.748	0.181	0.25	0.903	-0.41	0.806	0.094
upgma.14	0	554.646	0.536	1.001	-1	0.156	0.271	1354517.13	-504.263	0.087	0.25	0.944	-0.394	0.888	0.062
upgma.15	0	1105.132	0.451	1	-1	0.129	0.116	3291348.98	-556.815	0.041	0.25	0.958	-0.359	0.932	0.031
upgma.16	0	718469.797	0.014	38.855	-1	0.116	0.009	342211885	-590.161	0	0.25	0.991	-0.34	0.994	0
upgma.17	0	679164.134	0.014	0.856	-1	0.115	0.009	305213326	-589.562	0.103	0.243		-0.34	0.976	0.101
upgma.18	0	643502.067	0.014	0.89	-1	0.115	0.009	273925930	-588.961	0.125	0.236		-0.339	0.956	0.121
upgma.19	0	610392.811	0.014	0.736	-1	0.115	0.009	247106069	-588.366	0.18	0.229		-0.339	0.954	0.176
upgma.20	0	580974.819	0.014	0.736	-1	0.115	0.009	224205256	-587.76	0.179	0.224		-0.339	0.934	0.175
upgma.21	0	554358.464	0.014	0.736	-1	0.115	0.009	204415416	-587.153	0.178	0.218		-0.338	0.913	0.174
upgma.22	0	530275.736	0.014	0.736	-1	0.114	0.009	187212304	-586.547	0.178	0.213		-0.338	0.891	0.173
upgma.23	0	508409.671	0.014	0.766	-1	0.114	0.009	172176875	-585.939	0.177	0.209		-0.338	0.871	0.172
upgma.24	0	488306.783	0.014	0.744	-1	0.114	0.009	158922925	-585.33	0.188	0.204		-0.337	0.85	0.183
upgma.25	0	469716.65	0.044	0.731	-1	0.114	0.009	147170103	-584.721	0.202	0.2		-0.337	0.825	0.193
upgma.26	0	452693.017	0.043	0.731	-1	0.113	0.009	136739690	-584.121	0.201	0.196		-0.337	0.82	0.192
upgma.27	0	437057.906	0.041	0.731	-1	0.113	0.009	127436404	-583.511	0.2	0.192		-0.337	0.799	0.191
upgma.28	0	422452.628	0.04	0.727	-1	0.113	0.009	119069970	-582.9	0.202	0.189		-0.336	0.774	0.192
upgma.29	0	408890.068	0.054	0.721	-1	0.113	0.009	111533542	-582.298	0.202	0.186		-0.336	0.774	0.194
upgma.30	0	396260.314	0.067	0.721	-1	0.113	0.009	104723222	-581.706	0.203	0.183		-0.336	0.775	0.193
upgma.31	0	384587.822	0.065	0.721	-1	0.112	0.009	98565514.7	-581.103	0.202	0.18		-0.335	0.775	0.192
upgma.32	0	373617.973	0.063	0.733	-1	0.112	0.009	92954559.6	-580.49	0.201	0.177		-0.335	0.753	0.192

Tabella A.39: Ecoli Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.254	15.219	1.056	0.275	-0.636	0.064	0.688	0.097	-0.048	0.448	0.252	0.402	-0.178	0.469	1.171
upgma.3	0.111	136.71	0.903	0.113	-0.756	0.435	0.587	0.124	-0.145	0.281	0.349	0.447	-0.532	0.541	5.837
upgma.4	0.108	94.697	0.898	0.118	-0.762	0.437	0.583	0.073	-0.146	0.31	0.307	0.436	-0.537	0.518	5.728
upgma.5	0.101	81.041	0.881	0.118	-0.78	0.445	0.568	0.049	-0.15	0.467	0.282	0.364	-0.552	0.45	5.37
upgma.6	0.101	65.779	0.791	0.118	-0.78	0.445	0.568	0.053	-0.15	0.463	0.295		-0.552	0.451	5.324
upgma.7	0.1	57.699	0.791	0.118	-0.782	0.446	0.566	0.04	-0.151	0.451	0.284		-0.553	0.45	5.18
upgma.8	0.047	103.786	0.723	0.13	-0.911	0.411	0.46	0.048	-0.164	0.285	0.299		-0.597	0.521	5.389
upgma.9	0.047	94.32	0.72	0.13	-0.911	0.411	0.46	0.048	-0.164	0.264	0.283		-0.597	0.515	5.239
upgma.10	0.047	84.802	0.635	0.13	-0.911	0.41	0.46	0.039	-0.164	0.261	0.269		-0.597	0.516	5.186
upgma.11	0.045	80.444	0.622	0.13	-0.915	0.4	0.452	0.035	-0.163	0.427	0.258		-0.592	0.455	4.986
upgma.12	0.045	73.737	0.611	0.13	-0.915	0.4	0.452	0.032	-0.163	0.38	0.248		-0.592	0.464	4.946
upgma.13	0.045	68.375	0.613	0.13	-0.915	0.4	0.452	0.028	-0.163	0.405	0.238		-0.592	0.46	4.894
upgma.14	0.046	71.239	0.616	0.13	-0.916	0.378	0.442	0.026	-0.159	0.48	0.232		-0.576	0.429	4.47
upgma.15	0.044	75.222	0.558	0.139	-0.92	0.362	0.433	0.025	-0.157	0.435	0.23		-0.565	0.417	4.048
upgma.16	0.043	72.357	0.554	0.15	-0.923	0.356	0.428	0.022	-0.156	0.424	0.224		-0.562	0.396	3.945

Tabella A.40: S2 Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.213	3186.472	1.002	0.012	-0.559	0.385	0.631	7.9821E+10	-94445.057	0.348	0.439	0.416	-0.393	0.505	428.643
upgma.3	0.127	4194.6	0.837	0.014	-0.723	0.401	0.51	7.7714E+10	-118068.54	0.221	0.458	0.456	-0.494	0.524	262.025
upgma.4	0.087	4538.91	0.7	0.018	-0.805	0.363	0.44	1.05E+11	-121989.81	0.245	0.427	0.455	-0.511	0.522	188.409
upgma.5	0.078	4241.49	0.596	0.018	-0.825	0.338	0.416	8.724E+10	-120359.35	0.279	0.392	0.496	-0.502	0.523	159.652
upgma.6	0.059	4792.932	0.546	0.021	-0.868	0.29	0.363	7.9391E+10	-117021.23	0.254	0.371	0.489	-0.483	0.524	121.049
upgma.7	0.047	5103.81	0.52	0.024	-0.897	0.259	0.328	7.2853E+10	-114431.05	0.229	0.35	0.507	-0.469	0.548	98.403
upgma.8	0.042	5447.772	0.543	0.024	-0.913	0.24	0.306	8.35E+10	-112177.03	0.202	0.332	0.521	-0.457	0.547	81.25
upgma.9	0.038	5766.857	0.362	0.024	-0.922	0.228	0.292	9.8256E+10	-110790.39	0.181	0.317	0.556	-0.449	0.564	68.523
upgma.10	0.033	6336.629	0.347	0.026	-0.935	0.205	0.266	1.01E+11	-107391.28	0.206	0.303	0.576	-0.43	0.597	56.476
upgma.11	0.028	7127.446	0.35	0.026	-0.951	0.179	0.235	1.03E+11	-103134.91	0.195	0.291	0.573	-0.408	0.605	59.232
upgma.12	0.024	8002.696	0.35	0.026	-0.962	0.165	0.214	1.10E+11	-100746.1	0.162	0.281	0.588	-0.394	0.632	48.553
upgma.13	0.02	9126.835	0.326	0.026	-0.971	0.151	0.193	1.21E+11	-98304.185	0.143	0.271	0.598	-0.381	0.663	41.205
upgma.14	0.015	10645.308	0.324	0.026	-0.98	0.138	0.171	1.53E+11	-95681.544	0.128	0.263	0.616	-0.366	0.698	32.902
upgma.15	0.009	12485.469	0.309	0.02	-0.99	0.124	0.146	1.68E+11	-92489.323	0.13	0.255	0.61	-0.349	0.715	53.939
upgma.16	0.009	11690.25	0.314	0.02	-0.99	0.124	0.146	1.56E+11	-92471.537	0.13	0.247	0.618	-0.349	0.708	53.762
upgma.17	0.009	11271.044	0.318	0.02	-0.991	0.123	0.145	1.42E+11	-92223.786	0.149	0.239	0.602	-0.348	0.702	52.306
upgma.18	0.009	10646.774	0.295	0.02	-0.991	0.123	0.144	1.27E+11	-92189.408	0.159	0.233	0.607	-0.348	0.694	52.111
upgma.19	0.009	10265.995	0.305	0.023	-0.991	0.122	0.143	1.16E+11	-91958.882	0.175	0.226	0.579	-0.346	0.678	51.059
upgma.20	0.009	9854.098	0.315	0.023	-0.991	0.121	0.142	1.09E+11	-91814.361	0.177	0.221	0.573	-0.346	0.667	50.401
upgma.21	0.008	9553.462	0.323	0.025	-0.991	0.12	0.141	1.00E+11	-91585.689	0.178	0.215	0.56	-0.345	0.66	49.404
upgma.22	0.008	9117.164	0.53	0.025	-0.992	0.12	0.141	1.06E+11	-91559.179	0.181	0.21	0.558	-0.345	0.654	49.296
upgma.23	0.008	8781.874	0.511	0.025	-0.992	0.12	0.141	9.786E+10	-91464.002	0.18	0.206	0.56	-0.344	0.648	48.854
upgma.24	0.008	8557.201	0.514	0.025	-0.992	0.119	0.139	9.1328E+10	-91210.215	0.204	0.202	0.557	-0.343	0.639	47.969
upgma.25	0.008	8395.161	0.504	0.025	-0.992	0.118	0.138	8.575E+10	-90855.654	0.226	0.198	0.544	-0.341	0.634	46.875
upgma.26	0.008	8233.216	0.507	0.027	-0.992	0.117	0.137	8.0689E+10	-90473.756	0.226	0.194	0.539	-0.34	0.619	45.9
upgma.27	0.008	8158.994	0.464	0.028	-0.992	0.115	0.135	7.6592E+10	-89905.127	0.235	0.19	0.533	-0.337	0.612	44.558
upgma.28	0.008	8033.128	0.467	0.028	-0.992	0.114	0.134	7.2541E+10	-89447.824	0.287	0.187	0.526	-0.335	0.607	43.594
upgma.29	0.008	7844.19	0.469	0.028	-0.992	0.113	0.133	6.8273E+10	-89190.369	0.284	0.184	0.516	-0.334	0.597	43.053
upgma.30	0.008	7599.562	0.471	0.028	-0.992	0.113	0.133	6.4066E+10	-89186.661	0.283	0.181	0.518	-0.334	0.589	42.901

Tabella A.41: Wineqred Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.003	65.618	0.18	0.608	-1	0.005	0.213	14464.12	-9.428	0.03	0.069	0.864	-0.071	0.832	0.134
upgma.3	0.129	80.894	0.184	0.51	-0.943	0.04	0.383	7047.758	-11.566	0.16	0.193	0.686	-0.191	0.666	0.2
upgma.4	0.1	428.341	0.172	0.036	-0.787	0.335	0.474	6337.85	-19.398	0.181	0.214	0.614	-0.482	0.577	51.996
upgma.5	0.1	329.291	0.165	0.036	-0.787	0.335	0.474	5227.702	-19.399	0.197	0.196	0.642	-0.482	0.577	51.408
upgma.6	0.1	277.82	0.164	0.036	-0.786	0.337	0.475	3723.756	-19.397	0.327	0.182	0.602	-0.483	0.553	50.154
upgma.7	0.1	232.376	0.165	0.036	-0.786	0.337	0.475	2743.125	-19.397	0.327	0.169		-0.483	0.553	50.052
upgma.8	0.1	199.853	0.293	0.036	-0.786	0.337	0.475	2105.405	-19.397	0.348	0.16		-0.483	0.551	49.959
upgma.9	0.105	277.933	0.28	0.042	-0.731	0.418	0.487	2078.572	-18.668	0.332	0.164		-0.508	0.443	48.192
upgma.10	0.051	599.253	0.258	0.046	-0.877	0.407	0.377	3354.083	-18.673	0.218	0.176		-0.577	0.528	26.304
upgma.11	0.048	611.097	0.249	0.051	-0.884	0.405	0.372	3029.878	-18.721	0.198	0.169		-0.58	0.517	23.841
upgma.12	0.048	560.806	0.254	0.051	-0.885	0.405	0.372	2568.571	-18.721	0.215	0.163		-0.58	0.517	23.651

Tabella A.42: Breast Wisconsin Upgma

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
upgma.2	0.063	334.336	0.429	0.07	-0.931	0.122	0.255	1765979.89	-421.455	0.058	0.239	0.67	-0.332	0.749	9.484
upgma.3	0.063	176.034	0.202	0.075	-0.932	0.123	0.255	2535229.86	-421.672	0.082	0.202		-0.332	0.73	9.294
upgma.4	0.031	892.078	0.177	0.031	-0.929	0.467	0.224	5338766.1	-482.746	0.07	0.277		-0.647	0.741	36.687
upgma.5	0.031	683.672	0.181	0.031	-0.93	0.467	0.224	3466357.24	-482.808	0.085	0.253		-0.647	0.739	35.984
upgma.6	0.031	582.814	0.161	0.034	-0.93	0.467	0.223	2512294.43	-482.829	0.141	0.232		-0.647	0.708	34.078

Tabella A.43: Aggregation Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.131	692.981	0.918	0.159	-0.71	0.424	0.539	166.907	-4.958	0.259	0.449	0.464	-0.5	0.554	4.016
ward.3	0.065	988.52	0.657	0.135	-0.867	0.428	0.411	223.539	-5.785	0.165	0.483	0.532	-0.587	0.617	3.671
ward.4	0.043	1141.694	0.554	0.027	-0.919	0.402	0.358	290.417	-5.891	0.156	0.453	0.558	-0.595	0.632	99.961
ward.5	0.03	1208.002	0.531	0.03	-0.945	0.354	0.311	235.683	-5.7	0.206	0.415	0.533	-0.57	0.594	74.836
ward.6	0.021	1306.181	0.502	0.036	-0.962	0.337	0.289	238.817	-5.663	0.158	0.384	0.566	-0.564	0.608	57.388
ward.7	0.031	1376.768	0.502	0.039	-0.945	0.262	0.263	218.346	-4.977	0.36	0.359	0.509	-0.491	0.586	46.356
ward.8	0.031	1352.23	0.437	0.039	-0.944	0.231	0.247	193.542	-4.688	0.401	0.338	0.464	-0.46	0.543	40.856
ward.9	0.029	1355.101	0.429	0.039	-0.948	0.226	0.241	207.213	-4.657	0.353	0.321	0.526	-0.457	0.568	35.978
ward.10	0.031	1356.285	0.364	0.039	-0.945	0.201	0.231	189.577	-4.397	0.37	0.306	0.511	-0.43	0.574	32.155
ward.11	0.03	1375.863	0.368	0.039	-0.947	0.181	0.22	188.047	-4.195	0.33	0.293	0.479	-0.409	0.539	28.687
ward.12	0.028	1403.685	0.344	0.051	-0.951	0.174	0.213	174.92	-4.129	0.296	0.281	0.489	-0.401	0.549	25.681
ward.13	0.029	1404.592	0.349	0.051	-0.949	0.155	0.205	163.391	-3.905	0.425	0.271	0.464	-0.379	0.532	23.591
ward.14	0.03	1401.441	0.352	0.052	-0.948	0.143	0.2	152.583	-3.754	0.394	0.261	0.448	-0.363	0.528	21.87

Tabella A.44: Dim032 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.219	165.424	2.191	0.52	-0.313	0.322	0.855	17836.644	-32.599	1.327	0.266	0.215	-0.219	0.246	0.818
ward.3	0.141	177.947	1.983	0.564	-0.529	0.348	0.757	12777.782	-52.962	0.99	0.292	0.252	-0.357	0.273	0.705
ward.4	0.091	183.472	1.812	0.634	-0.705	0.338	0.686	12173.831	-64.014	0.805	0.295	0.327	-0.444	0.316	0.617
ward.5	0.058	201.01	1.273	0.634	-0.821	0.318	0.627	10102.232	-71.266	0.622	0.294	0.395	-0.485	0.376	0.531
ward.6	0.044	213.706	0.692	0.702	-0.884	0.304	0.592	14183.422	-74.74	0.544	0.292	0.495	-0.502	0.436	0.464
ward.7	0.035	233.219	0.692	0.709	-0.916	0.291	0.564	15031.179	-77.295	0.393	0.288	0.587	-0.505	0.495	0.4
ward.8	0.032	260.776	0.602	0.721	-0.937	0.233	0.47	13466.603	-82.996	0.424	0.283	0.59	-0.46	0.517	0.34
ward.9	0.024	298.274	0.602	0.779	-0.965	0.225	0.443	15945.307	-85.119	0.353	0.278	0.679	-0.462	0.582	0.284
ward.10	0.017	351.179	0.602	0.779	-0.976	0.203	0.388	17115.829	-88.278	0.288	0.273	0.727	-0.442	0.638	0.231
ward.11	0.007	426.362	1.019	0.824	-0.992	0.18	0.318	18951.894	-92.131	0.193	0.269	0.76	-0.422	0.69	0.182
ward.12	0.004	528.128	0.464	0.867	-0.996	0.168	0.277	24336.904	-94.019	0.149	0.265	0.808	-0.409	0.747	0.141
ward.13	0.002	694.258	0.487	0.881	-0.998	0.155	0.232	34941.088	-95.847	0.1	0.261	0.853	-0.394	0.808	0.103
ward.14	0.001	1033.514	0.491	0.901	-0.999	0.142	0.179	67900.516	-97.857	0.065	0.257	0.889	-0.377	0.863	0.066
ward.15	0	2063.91	0.381	0.942	-1	0.129	0.116	158833.618	-100.086	0.029	0.253	0.922	-0.359	0.918	0.032
ward.16	0	27667.061	0.063	4.035	-1	0.116	0.043	818251.881	-102.208	0.002	0.25	0.946	-0.34	0.965	0.003
ward.17	0	26258.766	0.063	0.899	-1	0.115	0.043	731385.152	-102.121	0.072	0.242		-0.34	0.952	0.079
ward.18	0	24903.25	0.078	0.46	-1	0.115	0.042	656560.678	-101.754	0.424	0.235		-0.338	0.939	0.301
ward.19	0	23675.85	0.093	0.46	-1	0.114	0.042	592008.949	-101.384	0.481	0.229		-0.337	0.912	0.299
ward.20	0	22568.813	0.094	0.46	-1	0.114	0.042	537905.604	-101.289	0.477	0.223		-0.337	0.897	0.297
ward.21	0	21567.771	0.094	0.476	-1	0.113	0.042	491004.362	-101.193	0.474	0.218		-0.337	0.884	0.295
ward.22	0	20663.399	0.095	0.476	-1	0.113	0.042	449323.448	-100.907	0.471	0.213		-0.336	0.859	0.293
ward.23	0	19840.603	0.094	0.431	-1	0.112	0.041	412794.71	-100.356	0.77	0.208		-0.334	0.822	0.374
ward.24	0	19085.989	0.092	0.432	-1	0.111	0.041	381495.764	-100.267	0.765	0.204		-0.334	0.823	0.371
ward.25	0	18392.536	0.093	0.432	-1	0.111	0.041	353093.845	-100.073	0.76	0.2		-0.333	0.798	0.369
ward.26	0	17755.472	0.093	0.432	-1	0.111	0.041	328441.68	-99.976	0.755	0.196		-0.333	0.783	0.366
ward.27	0	17168.612	0.093	0.411	-1	0.11	0.041	305791.564	-99.455	0.797	0.192		-0.331	0.778	0.426
ward.28	0	16626.371	0.093	0.411	-1	0.109	0.04	286126.642	-99.368	0.791	0.189		-0.331	0.779	0.423
ward.29	0.001	16121.986	0.092	0.406	-1	0.109	0.04	267867.095	-99.016	0.786	0.185		-0.33	0.776	0.432
ward.30	0.001	15653.044	0.092	0.406	-1	0.108	0.04	251340.615	-98.646	0.781	0.182		-0.328	0.77	0.429
ward.31	0.001	15210.376	0.088	0.418	-1	0.108	0.04	236834.94	-98.641	0.776	0.179		-0.328	0.771	0.427
ward.32	0.001	14796.832	0.088	0.418	-1	0.108	0.04	223738.175	-98.638	0.771	0.177		-0.328	0.774	0.424

Tabella A.45: Breast Tissue Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0	359.31	0.036	3.35	-1	0.037	0.047	1.3528E+10	-21880.943	0.003	0.318		-0.192	0.962	0.004
ward.3	0.002	827.826	0.031	0.521	-0.997	0.244	0.097	1.5818E+10	-15385.091	0.022	0.379		-0.492	0.862	0.262
ward.4	0.019	1671.179	0.028	0.084	-0.922	0.479	0.103	3.4924E+10	-8537.718	0.107	0.383		-0.651	0.758	4.471
ward.5	0.006	2155.252	0.02	0.084	-0.97	0.481	0.082	3.8542E+10	-8169.522	0.136	0.348		-0.678	0.743	2.597
ward.6	0.004	2984.653	0.012	0.173	-0.976	0.482	0.079	3.6464E+10	-8171.811	0.078	0.319		-0.682	0.747	1.493
ward.7	0.005	3341.338	0.012	0.075	-0.963	0.413	0.061	4.2484E+10	-6587.151	0.277	0.302		-0.625	0.7	5.819
ward.8	0.004	3770.416	0.011	0.089	-0.974	0.408	0.057	4.0088E+10	-6502.485	0.208	0.29		-0.626	0.681	4.381
ward.9	0.002	4186.008	0.009	0.089	-0.987	0.397	0.05	3.9605E+10	-6324.441	0.169	0.274		-0.624	0.695	3.42
ward.10	0.002	4334.794	0.009	0.121	-0.987	0.397	0.05	3.5452E+10	-6322.075	0.144	0.26		-0.624	0.693	2.907
ward.11	0.001	4498.262	0.009	0.121	-0.989	0.396	0.049	3.3363E+10	-6297.632	0.124	0.248		-0.624	0.696	2.496
ward.12	0.004	4697.328	0.009	0.039	-0.951	0.281	0.042	3.9737E+10	-4788.982	0.627	0.239		-0.51	0.669	20.464

Tabella A.46: Yeast Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.36	311.606	1.934	0.029	-0.277	0.319	0.823	0.018	-0.036	1.188	0.229	0.172	-0.196	0.269	45.834
ward.3	0.257	320.661	1.677	0.03	-0.433	0.338	0.729	0.016	-0.054	0.851	0.247	0.203	-0.297	0.313	38.709
ward.4	0.222	291.898	1.36	0.03	-0.493	0.333	0.694	0.023	-0.06	1.452	0.239	0.209	-0.329	0.267	34.85
ward.5	0.174	288.232	1.36	0.03	-0.59	0.324	0.642	0.026	-0.067	1.297	0.241	0.221	-0.377	0.285	31.172
ward.6	0.151	299.916	1.29	0.03	-0.619	0.327	0.615	0.037	-0.072	1.13	0.261	0.27	-0.393	0.294	27.534
ward.7	0.138	279.141	0.941	0.03	-0.643	0.328	0.6	0.034	-0.075	1.096	0.275	0.265	-0.406	0.297	25.994
ward.8	0.146	262.314	0.941	0.03	-0.616	0.249	0.604	0.027	-0.063	1.943	0.261	0.219	-0.342	0.239	24.719
ward.9	0.143	248.267	0.941	0.03	-0.623	0.238	0.599	0.023	-0.061	1.57	0.252	0.223	-0.338	0.247	23.639
ward.10	0.135	236.375	0.904	0.03	-0.645	0.225	0.587	0.019	-0.061	1.66	0.242	0.209	-0.337	0.245	22.704
ward.11	0.123	227.241	0.904	0.039	-0.673	0.212	0.568	0.017	-0.061	1.595	0.233	0.192	-0.339	0.213	21.816
ward.12	0.119	219.251	0.878	0.039	-0.685	0.197	0.56	0.015	-0.06	1.385	0.225	0.18	-0.331	0.216	21.024
ward.13	0.117	212.184	0.377	0.042	-0.688	0.197	0.557	0.02	-0.06	1.338	0.22	0.227	-0.333	0.217	20.312
ward.14	0.112	206.242	0.374	0.042	-0.7	0.197	0.55	0.018	-0.061	1.294	0.213	0.233	-0.337	0.222	19.643
ward.15	0.112	201.805	0.374	0.042	-0.698	0.153	0.543	0.016	-0.054	1.803	0.208	0.215	-0.297	0.219	18.976
ward.16	0.109	196.4	0.375	0.042	-0.707	0.147	0.536	0.014	-0.053	1.467	0.204	0.206	-0.294	0.224	18.448
ward.17	0.105	191.749	0.375	0.042	-0.721	0.143	0.528	0.013	-0.053	1.427	0.199	0.199	-0.294	0.212	17.944
ward.18	0.101	187.901	0.375	0.042	-0.73	0.142	0.522	0.012	-0.053	1.387	0.195	0.2	-0.296	0.212	17.449
ward.19	0.098	184.594	0.376	0.042	-0.741	0.138	0.515	0.011	-0.053	1.349	0.19	0.197	-0.294	0.21	16.974
ward.20	0.097	181.536	0.369	0.042	-0.743	0.123	0.511	0.01	-0.05	1.324	0.186	0.189	-0.279	0.212	16.529

Tabella A.47: Dim064 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.166	121.151	2.638	0.721	-0.389	0.34	0.863	23736.09	-41.715	1.812	0.222	0.178	-0.272	0.197	0.597
ward.3	0.099	136.77	2.264	0.758	-0.617	0.36	0.77	16324.108	-66.83	1.276	0.259	0.221	-0.412	0.237	0.527
ward.4	0.094	138.13	1.513	0.758	-0.623	0.322	0.731	15230.499	-73.392	1.053	0.265	0.31	-0.392	0.29	0.475
ward.5	0.079	145.797	1.499	0.777	-0.695	0.296	0.682	14276.38	-81.246	0.826	0.265	0.373	-0.411	0.334	0.425
ward.6	0.061	157.239	1.485	0.826	-0.772	0.26	0.612	12292.371	-90.859	0.606	0.266	0.415	-0.418	0.377	0.377
ward.7	0.05	171.765	0.628	0.826	-0.818	0.248	0.577	14452.185	-95.317	0.533	0.265	0.515	-0.427	0.431	0.332
ward.8	0.032	192.038	0.618	0.826	-0.879	0.215	0.49	15950.426	-105.017	0.43	0.263	0.545	-0.421	0.47	0.288
ward.9	0.02	219.469	0.469	0.858	-0.934	0.199	0.43	17467.78	-111.18	0.346	0.262	0.61	-0.423	0.534	0.245
ward.10	0.012	255.936	0.476	0.88	-0.965	0.19	0.392	22863.083	-114.874	0.273	0.261	0.693	-0.424	0.6	0.204
ward.11	0.005	309.17	0.45	0.911	-0.986	0.18	0.35	27387.061	-118.772	0.22	0.26	0.76	-0.42	0.668	0.165
ward.12	0.002	383.697	0.454	0.94	-0.995	0.168	0.305	35398.203	-122.452	0.16	0.258	0.811	-0.408	0.736	0.129
ward.13	0.001	503.251	0.391	0.949	-0.997	0.155	0.254	50968.002	-126.207	0.119	0.256	0.858	-0.393	0.799	0.096
ward.14	0	748.91	0.38	0.968	-0.999	0.142	0.193	90766.294	-130.525	0.075	0.253	0.9	-0.377	0.86	0.063
ward.15	0	1488.164	0.387	0.986	-1	0.129	0.119	238833.094	-135.418	0.032	0.252	0.936	-0.359	0.923	0.031
ward.16	0	54451.234	0.038	5.82	-1	0.116	0.03	2664252.09	-140.874	0.001	0.25	0.966	-0.34	0.978	0.001
ward.17	0	52092.055	0.038	1.013	-1	0.115	0.03	2389220.81	-140.753	0.045	0.242		-0.34	0.971	0.046
ward.18	0	49643.355	0.038	0.869	-1	0.115	0.029	2151346.75	-140.625	0.071	0.236		-0.339	0.96	0.074
ward.19	0	47417.285	0.038	0.811	-1	0.115	0.029	1948154.42	-140.495	0.078	0.229		-0.339	0.948	0.084
ward.20	0	45381.704	0.038	0.797	-1	0.115	0.029	1773525.44	-140.366	0.086	0.223		-0.339	0.946	0.093
ward.21	0	43445.406	0.044	0.493	-1	0.113	0.029	1616866.54	-139.392	0.785	0.218		-0.336	0.906	0.272
ward.22	0	41689.636	0.044	0.502	-1	0.113	0.029	1480901.18	-139.159	0.778	0.213		-0.336	0.907	0.269
ward.23	0	40090.505	0.044	0.513	-1	0.113	0.029	1362008.71	-139.127	0.53	0.208		-0.336	0.911	0.267
ward.24	0	38584.365	0.044	0.513	-1	0.112	0.028	1256701.89	-138.9	0.535	0.204		-0.335	0.911	0.265
ward.25	0	37199.868	0.044	0.513	-1	0.112	0.028	1165916.97	-138.765	0.531	0.2		-0.335	0.907	0.263
ward.26	0	35925.791	0.047	0.513	-1	0.112	0.028	1082590.19	-138.492	0.528	0.196		-0.334	0.882	0.261
ward.27	0	34751.325	0.045	0.515	-1	0.112	0.028	1008452.68	-138.479	0.374	0.192		-0.334	0.885	0.26
ward.28	0	33663.144	0.045	0.515	-1	0.111	0.028	942113.41	-138.213	0.371	0.189		-0.334	0.885	0.258
ward.29	0	32650.867	0.045	0.545	-1	0.111	0.028	884017.136	-138.075	0.368	0.186		-0.333	0.866	0.256
ward.30	0	31700.812	0.045	0.545	-1	0.111	0.028	829952.71	-137.808	0.366	0.182		-0.333	0.858	0.254
ward.31	0	30809.788	0.046	0.498	-1	0.11	0.028	784634.683	-137.275	0.571	0.179		-0.332	0.846	0.302
ward.32	0	29978.068	0.048	0.498	-1	0.11	0.028	739473.303	-137.003	0.567	0.177		-0.331	0.836	0.3

Tabella A.48: Dim128 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.106	103.326	3.045	0.81	-0.503	0.374	0.854	37284.288	-63.525	2.318	0.209	0.151	-0.355	0.164	0.534
ward.3	0.11	106.384	2.155	0.81	-0.456	0.338	0.809	31703.932	-79.683	1.828	0.235	0.225	-0.31	0.207	0.487
ward.4	0.067	114.692	1.764	0.81	-0.658	0.351	0.762	32539.563	-95.331	1.342	0.248	0.311	-0.428	0.257	0.44
ward.5	0.06	123.331	0.759	0.81	-0.695	0.353	0.754	55315.486	-97.734	1.209	0.252	0.544	-0.449	0.334	0.396
ward.6	0.05	135.288	0.691	0.81	-0.738	0.338	0.728	43566.8	-104.396	0.978	0.255	0.579	-0.46	0.382	0.353
ward.7	0.034	151.435	0.628	0.881	-0.826	0.259	0.604	36080.71	-128.932	0.779	0.256	0.56	-0.44	0.419	0.321
ward.8	0.028	169.934	0.694	0.881	-0.857	0.244	0.568	35302.966	-135.263	0.679	0.257	0.634	-0.439	0.478	0.28
ward.9	0.022	195.559	0.699	0.883	-0.897	0.228	0.524	35493.293	-142.581	0.475	0.257	0.689	-0.44	0.546	0.239
ward.10	0.012	230.081	0.551	0.883	-0.949	0.188	0.403	33866.051	-160.082	0.317	0.256	0.695	-0.417	0.591	0.2
ward.11	0.005	280.43	0.499	0.931	-0.991	0.18	0.359	41849.446	-166.278	0.256	0.256	0.764	-0.422	0.658	0.162
ward.12	0.001	351.357	0.494	0.969	-0.997	0.168	0.311	55066.831	-172.236	0.196	0.255	0.82	-0.409	0.727	0.126
ward.13	0.001	464.651	0.479	0.969	-0.998	0.155	0.257	81325.673	-178.288	0.13	0.254	0.869	-0.394	0.796	0.093
ward.14	0	698.166	0.423	0.998	-1	0.143	0.193	148720.191	-185.202	0.085	0.253	0.91	-0.378	0.861	0.061
ward.15	0	1358.692	0.432	0.982	-0.999	0.129	0.118	405285.107	-192.576	0.033	0.251	0.945	-0.359	0.928	0.032
ward.16	0	86413.57	0.049	7.48	-1	0.116	0.023	7119600.46	-201.464	0	0.25	0.975	-0.34	0.983	0
ward.17	0	82851.132	0.049	0.806	-1	0.115	0.023	6418142.2	-201.285	0.041	0.242		-0.34	0.975	0.042
ward.18	0	79598.228	0.048	0.922	-1	0.115	0.023	5792281.28	-201.106	0.044	0.236		-0.339	0.962	0.046
ward.19	0	76337.792	0.048	0.804	-1	0.115	0.023	5259381.52	-200.919	0.058	0.229		-0.339	0.956	0.06
ward.20	0	73361.963	0.047	0.838	-1	0.115	0.023	4793951.51	-200.736	0.062	0.224		-0.339	0.954	0.065
ward.21	0	70510.425	0.052	0.762	-1	0.115	0.023	4387175.65	-200.546	0.076	0.218		-0.338	0.937	0.078
ward.22	0	67947.381	0.051	0.761	-1	0.114	0.023	4034518.83	-200.355	0.075	0.213		-0.338	0.927	0.077
ward.23	0	65597.72	0.051	0.62	-1	0.114	0.022	3713831.02	-199.973	0.153	0.208		-0.337	0.907	0.115
ward.24	0	63457.209	0.05	0.62	-1	0.114	0.022	3441511.64	-199.781	0.151	0.204		-0.337	0.892	0.113
ward.25	0	61455.355	0.048	0.62	-1	0.113	0.022	3198766.73	-199.595	0.149	0.2		-0.337	0.89	0.112
ward.26	0	59576.659	0.047	0.62	-1	0.113	0.022	2984141.51	-199.404	0.147	0.196		-0.337	0.886	0.111
ward.27	0	57795.402	0.06	0.671	-1	0.113	0.022	2780906.49	-199.038	0.191	0.192		-0.336	0.879	0.111
ward.28	0	56148.495	0.059	0.7	-1	0.113	0.022	2605673.05	-198.859	0.19	0.189		-0.336	0.88	0.11
ward.29	0	54585.584	0.043	0.726	-1	0.113	0.022	2454415.58	-198.856	0.143	0.186		-0.336	0.886	0.109
ward.30	0	53106.461	0.043	0.708	-1	0.112	0.022	2311391.78	-198.66	0.142	0.183		-0.335	0.884	0.113
ward.31	0	51729.121	0.041	0.767	-1	0.112	0.022	2186623.07	-198.657	0.11	0.18		-0.335	0.888	0.112
ward.32	0	50413.382	0.04	0.728	-1	0.112	0.022	2066439.28	-198.458	0.119	0.177		-0.335	0.876	0.123

Tabella A.49: Flame Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.241	122.782	1.185	0.037	-0.499	0.373	0.664	12.83	-1.217	0.446	0.412	0.354	-0.352	0.446	50.469
ward.3	0.133	176.904	0.807	0.045	-0.721	0.387	0.513	13.756	-1.65	0.237	0.447	0.397	-0.483	0.495	30.69
ward.4	0.103	200.909	0.807	0.051	-0.788	0.358	0.458	18.424	-1.7	0.209	0.424	0.407	-0.499	0.497	21.527
ward.5	0.094	199.617	0.724	0.055	-0.811	0.293	0.417	15.79	-1.583	0.4	0.393	0.363	-0.461	0.482	18.902
ward.6	0.077	213.208	0.502	0.055	-0.851	0.27	0.381	19.572	-1.58	0.317	0.369	0.391	-0.46	0.483	14.962

Tabella A.50: Dim256 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.094	98.978	3.187	0.897	-0.422	0.356	0.859	70315.367	-87.988	2.541	0.211	0.142	-0.299	0.153	0.479
ward.3	0.097	100.758	2.494	0.86	-0.42	0.323	0.811	52491.879	-112.164	1.776	0.233	0.205	-0.283	0.202	0.479
ward.4	0.06	107.92	1.77	0.873	-0.62	0.321	0.752	58446.783	-137.065	1.415	0.244	0.294	-0.39	0.244	0.435
ward.5	0.046	115.869	1.617	0.878	-0.711	0.299	0.703	51319.024	-154.091	1.247	0.249	0.361	-0.421	0.286	0.394
ward.6	0.035	126.586	1.508	0.897	-0.782	0.262	0.632	40387.81	-175.102	0.878	0.252	0.407	-0.424	0.338	0.353
ward.7	0.032	139.521	1.505	0.897	-0.81	0.228	0.563	34294.217	-192.24	0.659	0.253	0.455	-0.406	0.389	0.314
ward.8	0.026	155.839	1.454	0.897	-0.853	0.212	0.514	44226.723	-203.853	0.51	0.253	0.542	-0.408	0.448	0.276
ward.9	0.018	178.051	0.563	0.906	-0.913	0.197	0.451	54812.634	-217.838	0.382	0.254	0.61	-0.414	0.515	0.239
ward.10	0.012	208.582	0.562	0.933	-0.962	0.19	0.413	62104.411	-226.164	0.322	0.254	0.696	-0.423	0.583	0.201
ward.11	0.009	249.888	0.446	0.951	-0.98	0.179	0.37	74379.01	-234.652	0.265	0.254	0.766	-0.417	0.652	0.165
ward.12	0.007	310.355	0.454	0.957	-0.985	0.167	0.322	106243.821	-243.561	0.208	0.253	0.823	-0.404	0.723	0.131
ward.13	0.005	412.614	0.459	0.96	-0.99	0.155	0.265	158107.952	-253.522	0.154	0.253	0.873	-0.391	0.789	0.097
ward.14	0.002	621.767	0.458	0.97	-0.995	0.142	0.197	294901.057	-264.849	0.095	0.252	0.915	-0.376	0.86	0.064
ward.15	0	1271.989	0.277	1.006	-1	0.129	0.115	898615.401	-277.806	0.036	0.251	0.952	-0.359	0.93	0.031
ward.16	0	203865.162	0.022	14.85	-1	0.116	0.016	30072605.6	-291.899	0	0.25	0.983	-0.34	0.989	0
ward.17	0	193303.567	0.022	0.873	-1	0.115	0.016	26864475.7	-291.614	0.077	0.242		-0.34	0.973	0.079
ward.18	0	183506.643	0.022	0.766	-1	0.115	0.016	24139720	-291.33	0.1	0.236		-0.339	0.972	0.101
ward.19	0	174639.549	0.029	0.72	-1	0.115	0.016	21759696.1	-290.759	0.218	0.229		-0.339	0.947	0.118
ward.20	0	166679.873	0.029	0.72	-1	0.115	0.016	19772653.9	-290.467	0.217	0.224		-0.338	0.932	0.117
ward.21	0	159465.824	0.022	0.789	-1	0.115	0.016	18101638.9	-290.463	0.114	0.218		-0.338	0.939	0.116
ward.22	0	152787.224	0.029	0.659	-1	0.114	0.016	16556649.1	-289.636	0.397	0.213		-0.338	0.928	0.17
ward.23	0	146707.071	0.03	0.659	-1	0.114	0.016	15241860.4	-289.343	0.394	0.208		-0.337	0.908	0.169
ward.24	0	141160.173	0.03	0.659	-1	0.113	0.016	14057065.3	-288.761	0.391	0.204		-0.337	0.885	0.168
ward.25	0	136027.722	0.029	0.679	-1	0.113	0.016	13032413.5	-288.49	0.389	0.2		-0.336	0.886	0.167
ward.26	0	131292.089	0.029	0.687	-1	0.113	0.016	12120585.1	-288.203	0.386	0.196		-0.336	0.883	0.166
ward.27	0	126911.574	0.029	0.645	-1	0.112	0.015	11281631.9	-287.424	0.44	0.192		-0.335	0.882	0.187
ward.28	0	122847.72	0.041	0.645	-1	0.112	0.015	10592888.6	-287.127	0.438	0.189		-0.335	0.865	0.186
ward.29	0	119080.732	0.04	0.645	-1	0.112	0.015	9932748.82	-286.835	0.435	0.186		-0.335	0.865	0.184
ward.30	0	115570.734	0.04	0.65	-1	0.112	0.015	9335680.26	-286.537	0.432	0.183		-0.334	0.844	0.183
ward.31	0	112301.603	0.038	0.659	-1	0.112	0.015	8795427.86	-286.528	0.43	0.18		-0.334	0.845	0.182
ward.32	0	109243.919	0.037	0.674	-1	0.112	0.015	8322774.36	-286.523	0.427	0.177		-0.334	0.851	0.181

Tabella A.51: Ecoli Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.145	217.793	0.978	0.102	-0.684	0.411	0.621	0.146	-0.13	0.324	0.341	0.429	-0.478	0.537	6.434
ward.3	0.096	229.285	0.866	0.087	-0.809	0.41	0.519	0.121	-0.15	0.372	0.362	0.388	-0.545	0.526	7.593
ward.4	0.075	188.169	0.95	0.087	-0.854	0.412	0.494	0.089	-0.156	0.327	0.369	0.366	-0.57	0.545	6.684
ward.5	0.083	170.031	0.959	0.07	-0.841	0.378	0.487	0.066	-0.149	0.56	0.338	0.317	-0.539	0.497	9.149
ward.6	0.109	155.128	0.947	0.08	-0.797	0.282	0.483	0.063	-0.124	1.117	0.314	0.249	-0.446	0.385	9.403
ward.7	0.1	144.914	1.168	0.08	-0.814	0.261	0.465	0.05	-0.122	1.027	0.3	0.246	-0.436	0.371	8.648
ward.8	0.084	137.641	1.179	0.082	-0.845	0.255	0.443	0.041	-0.124	1.101	0.284	0.231	-0.444	0.377	8.001
ward.9	0.088	131.28	1.112	0.081	-0.838	0.221	0.437	0.039	-0.115	1.066	0.27	0.226	-0.411	0.377	7.618
ward.10	0.083	127.399	1.009	0.081	-0.849	0.202	0.424	0.033	-0.112	0.994	0.26	0.215	-0.397	0.368	7.103
ward.11	0.081	122.913	1.018	0.089	-0.852	0.202	0.422	0.035	-0.112	0.939	0.251	0.25	-0.398	0.373	6.71
ward.12	0.082	119.701	1.026	0.089	-0.85	0.181	0.416	0.031	-0.106	0.814	0.241	0.243	-0.376	0.376	6.336
ward.13	0.081	114.935	1.033	0.107	-0.852	0.181	0.414	0.027	-0.107	0.782	0.233	0.265	-0.377	0.381	6.089
ward.14	0.081	110.836	1.038	0.107	-0.854	0.158	0.405	0.025	-0.1	1.103	0.226	0.243	-0.353	0.335	5.861
ward.15	0.077	107.484	1.043	0.119	-0.862	0.149	0.396	0.023	-0.098	1.062	0.219	0.234	-0.345	0.328	5.641
ward.16	0.074	104.905	1.048	0.119	-0.867	0.142	0.389	0.021	-0.097	0.885	0.213	0.234	-0.338	0.334	5.422

Tabella A.52: Iris Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.023	501.925	0.384	0.339	-0.958	0.484	0.324	20.4	-1.365	0.066	0.56	0.722	-0.674	0.774	0.384
ward.3	0.033	556.841	0.613	0.113	-0.915	0.434	0.274	25.367	-1.174	0.161	0.501	0.56	-0.616	0.667	5.292
ward.4	0.03	513.772	0.76	0.124	-0.908	0.365	0.251	19.025	-1.033	0.296	0.445	0.458	-0.562	0.607	4.364
ward.5	0.023	487.07	0.602	0.124	-0.924	0.343	0.235	18.842	-0.999	0.297	0.403	0.436	-0.552	0.606	3.493
ward.6	0.029	465.732	0.528	0.131	-0.895	0.276	0.231	16.667	-0.864	0.434	0.377	0.37	-0.483	0.511	5.287

Tabella A.53: Dim512 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.091	88.709	3.266	0.879	-0.351	0.336	0.874	132088.277	-110.515	2.7	0.2	0.142	-0.247	0.146	0.497
ward.3	0.066	96.155	2.648	0.879	-0.514	0.338	0.802	94774.858	-163.793	1.807	0.229	0.198	-0.343	0.195	0.455
ward.4	0.05	102.482	1.763	0.901	-0.617	0.321	0.757	102198.817	-189.596	1.527	0.239	0.291	-0.388	0.24	0.415
ward.5	0.036	111.266	0.755	0.91	-0.732	0.317	0.724	153625.103	-206.874	1.236	0.245	0.442	-0.443	0.301	0.376
ward.6	0.025	121.742	0.728	0.924	-0.822	0.286	0.662	118763.693	-234.761	1.026	0.249	0.473	-0.46	0.345	0.338
ward.7	0.021	133.408	0.728	0.932	-0.848	0.252	0.603	102891.758	-256.956	0.803	0.249	0.509	-0.443	0.39	0.302
ward.8	0.016	148.671	0.622	0.938	-0.873	0.215	0.517	88817.331	-285.945	0.53	0.25	0.538	-0.418	0.442	0.267
ward.9	0.012	169.175	0.625	0.938	-0.906	0.207	0.487	115970.286	-295.645	0.46	0.251	0.643	-0.423	0.515	0.232
ward.10	0.009	197.544	0.627	0.949	-0.943	0.2	0.454	131297.398	-306.257	0.39	0.252	0.727	-0.428	0.586	0.196
ward.11	0.006	237.338	0.558	0.949	-0.963	0.178	0.375	143522.885	-328.725	0.285	0.252	0.765	-0.41	0.647	0.162
ward.12	0.003	299.115	0.555	0.973	-0.993	0.168	0.324	193258.562	-342.707	0.224	0.252	0.825	-0.408	0.719	0.127
ward.13	0.002	398.902	0.511	0.98	-0.997	0.155	0.266	289625.329	-357.396	0.155	0.252	0.875	-0.393	0.79	0.094
ward.14	0	601.426	0.514	0.997	-1	0.143	0.197	545722.147	-373.96	0.101	0.252	0.918	-0.378	0.859	0.062
ward.15	0	1201.176	0.48	1.007	-1	0.129	0.115	1680087.95	-392.397	0.037	0.251	0.954	-0.359	0.931	0.031
ward.16	0	330337.86	0.017	20.008	-1	0.116	0.013	87532562.7	-413.753	0	0.25	0.987	-0.34	0.991	0
ward.17	0	313156.831	0.017	0.75	-1	0.115	0.013	78175047.4	-413.342	0.079	0.243		-0.34	0.978	0.08
ward.18	0	297818.927	0.017	0.758	-1	0.115	0.013	70280382.2	-412.932	0.084	0.236		-0.339	0.959	0.084
ward.19	0	283887.604	0.018	0.748	-1	0.115	0.013	63550627	-412.518	0.093	0.229		-0.339	0.944	0.095
ward.20	0	271247.252	0.018	0.722	-1	0.115	0.013	57774167.5	-412.109	0.101	0.224		-0.339	0.945	0.1
ward.21	0	259815.383	0.018	0.769	-1	0.115	0.012	52773568	-411.701	0.104	0.218		-0.338	0.943	0.104
ward.22	0	249269.162	0.017	0.739	-1	0.114	0.012	48406772.4	-411.298	0.115	0.213		-0.338	0.943	0.116
ward.23	0	239579.648	0.017	0.762	-1	0.114	0.012	44576448	-410.9	0.122	0.208		-0.338	0.943	0.122
ward.24	0	230535.547	0.021	0.681	-1	0.113	0.012	41090544.9	-409.402	0.521	0.204		-0.337	0.93	0.164
ward.25	0	222219.006	0.022	0.681	-1	0.113	0.012	38012892.9	-408.678	0.523	0.2		-0.336	0.931	0.163
ward.26	0	214514.386	0.021	0.681	-1	0.113	0.012	35353092.8	-408.256	0.519	0.196		-0.336	0.912	0.162
ward.27	0	207407.52	0.022	0.699	-1	0.113	0.012	33052142.2	-408.249	0.516	0.192		-0.336	0.913	0.161
ward.28	0	200834.671	0.021	0.699	-1	0.113	0.012	30914257.6	-407.826	0.513	0.189		-0.336	0.895	0.16
ward.29	0	194733.143	0.021	0.699	-1	0.112	0.012	28987184.7	-407.402	0.509	0.186		-0.335	0.874	0.159
ward.30	0	189036.387	0.021	0.699	-1	0.112	0.012	27277891.7	-406.985	0.506	0.183		-0.335	0.869	0.158
ward.31	0	183708.014	0.021	0.713	-1	0.112	0.012	25694789.7	-406.965	0.384	0.18		-0.335	0.872	0.157
ward.32	0	178691.238	0.02	0.713	-1	0.112	0.012	24254402.8	-406.951	0.279	0.177		-0.335	0.876	0.156

Tabella A.54: Jain Toy Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.124	468.062	0.8	0.032	-0.723	0.43	0.485	185.313	-5.121	0.19	0.532	0.487	-0.511	0.582	83.857
ward.3	0.099	527.361	0.579	0.032	-0.783	0.402	0.411	231.245	-5.111	0.208	0.487	0.477	-0.526	0.576	49.253
ward.4	0.074	517.049	0.529	0.035	-0.84	0.357	0.349	253.907	-5.05	0.301	0.446	0.425	-0.524	0.547	36.446
ward.5	0.062	563.525	0.439	0.021	-0.867	0.302	0.304	259.066	-4.756	0.236	0.414	0.439	-0.493	0.544	72.281
ward.6	0.039	713.252	0.388	0.032	-0.919	0.288	0.26	370.576	-4.845	0.157	0.386	0.484	-0.504	0.575	48.055

Tabella A.55: Pathbased Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.128	318.422	0.798	0.051	-0.741	0.435	0.505	134.609	-4.175	0.232	0.446	0.498	-0.524	0.625	23.138
ward.3	0.057	336.432	0.663	0.061	-0.881	0.45	0.403	115.328	-4.852	0.168	0.466	0.478	-0.61	0.637	14.657
ward.4	0.075	294.906	0.787	0.059	-0.845	0.388	0.387	121.529	-4.36	0.471	0.428	0.381	-0.548	0.546	14.601
ward.5	0.068	291.101	0.723	0.073	-0.86	0.385	0.376	119.334	-4.396	0.38	0.393	0.431	-0.553	0.537	11.773
ward.6	0.078	291.541	1.168	0.035	-0.839	0.304	0.354	98.974	-3.842	0.492	0.367	0.383	-0.482	0.471	52.709

Tabella A.56: Dim1024 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.06	85.607	3.455	0.932	-0.4	0.35	0.868	240955.072	-163.677	2.985	0.196	0.131	-0.282	0.134	0.466
ward.3	0.058	90.272	2.986	0.932	-0.396	0.33	0.832	235112.993	-202.612	2.144	0.221	0.225	-0.272	0.186	0.429
ward.4	0.04	97.489	2.807	0.932	-0.548	0.318	0.777	238346.417	-251.013	1.782	0.235	0.305	-0.351	0.225	0.393
ward.5	0.028	105.387	0.796	0.932	-0.681	0.314	0.741	305623.056	-278.585	1.514	0.241	0.448	-0.416	0.286	0.357
ward.6	0.019	115.517	0.796	0.958	-0.767	0.285	0.683	235676.308	-316.815	1.043	0.245	0.479	-0.436	0.334	0.322
ward.7	0.008	127.701	0.648	0.967	-0.893	0.249	0.59	192787.282	-368.76	0.738	0.247	0.499	-0.458	0.389	0.288
ward.8	0.006	142.683	0.578	0.969	-0.928	0.243	0.569	238294.596	-379.754	0.653	0.248	0.623	-0.465	0.464	0.255
ward.9	0.003	162.543	0.589	0.984	-0.979	0.227	0.519	245239.925	-403.862	0.567	0.249	0.682	-0.469	0.522	0.221
ward.10	0.002	188.025	0.599	0.985	-0.986	0.216	0.491	279156.769	-416.795	0.485	0.249	0.761	-0.46	0.595	0.189
ward.11	0.002	224.097	0.595	0.988	-0.992	0.193	0.422	306954.578	-446.302	0.337	0.25	0.797	-0.436	0.656	0.157
ward.12	0.001	278.787	0.602	0.991	-0.996	0.181	0.379	411809.677	-463.532	0.269	0.25	0.854	-0.424	0.729	0.125
ward.13	0	370.625	0.608	0.994	-0.998	0.168	0.329	645479.778	-482.748	0.181	0.25	0.903	-0.41	0.806	0.094
ward.14	0	554.646	0.536	1.001	-1	0.156	0.271	1354517.13	-504.263	0.087	0.25	0.944	-0.394	0.888	0.062
ward.15	0	1105.132	0.451	1	-1	0.129	0.116	3291348.98	-556.815	0.041	0.25	0.958	-0.359	0.932	0.031
ward.16	0	718469.797	0.014	38.855	-1	0.116	0.009	342211885	-590.161	0	0.25	0.991	-0.34	0.994	0
ward.17	0	679164.134	0.014	0.856	-1	0.115	0.009	305213326	-589.562	0.103	0.243		-0.34	0.976	0.101
ward.18	0	643502.067	0.014	0.89	-1	0.115	0.009	273925930	-588.961	0.125	0.236		-0.339	0.956	0.121
ward.19	0	610580.132	0.014	0.761	-1	0.115	0.009	247161159	-588.356	0.17	0.229		-0.339	0.937	0.164
ward.20	0	580977.37	0.014	0.741	-1	0.115	0.009	224216804	-587.75	0.179	0.224		-0.339	0.915	0.172
ward.21	0	554358.464	0.014	0.736	-1	0.115	0.009	204415416	-587.153	0.178	0.218		-0.338	0.913	0.174
ward.22	0	530279.247	0.014	0.754	-1	0.114	0.009	187219829	-586.546	0.178	0.213		-0.338	0.893	0.173
ward.23	0	508409.671	0.014	0.766	-1	0.114	0.009	172176875	-585.939	0.177	0.209		-0.338	0.871	0.172
ward.24	0	488306.783	0.014	0.744	-1	0.114	0.009	158922925	-585.33	0.188	0.204		-0.337	0.85	0.183
ward.25	0	469848.152	0.014	0.743	-1	0.114	0.009	147196742	-584.721	0.191	0.2		-0.337	0.829	0.182
ward.26	0	452820.377	0.014	0.73	-1	0.113	0.009	136764495	-584.121	0.195	0.196		-0.337	0.823	0.188
ward.27	0	437110.784	0.014	0.729	-1	0.113	0.009	127448478	-583.52	0.195	0.192		-0.337	0.824	0.187
ward.28	0	422575.362	0.128	0.721	-1	0.113	0.009	118866836	-582.317	0.383	0.189		-0.336	0.792	0.195
ward.29	0	409082.944	0.125	0.721	-1	0.113	0.009	111361242	-581.705	0.381	0.186		-0.336	0.773	0.194
ward.30	0	396501.361	0.058	0.721	-1	0.112	0.009	104569405	-581.112	0.38	0.183		-0.335	0.774	0.193
ward.31	0	384766.991	0.056	0.721	-1	0.112	0.009	98412131.4	-580.508	0.378	0.18		-0.335	0.773	0.192
ward.32	0	373792.879	0.055	0.733	-1	0.112	0.009	92809554.4	-579.895	0.376	0.177		-0.335	0.751	0.192

Tabella A.57: Glass Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.068	114.502	1.177	0.102	-0.903	0.444	0.356	5.523	-1.517	0.341	0.297	0.406	-0.617	0.624	2.695
ward.3	0.032	124.831	0.98	0.13	-0.949	0.467	0.317	5.714	-1.647	0.217	0.334	0.429	-0.658	0.665	1.901
ward.4	0.023	116.568	0.769	0.134	-0.963	0.473	0.308	8.96	-1.678	0.149	0.328	0.404	-0.669	0.694	1.557
ward.5	0.034	115.513	0.721	0.065	-0.931	0.468	0.278	7.507	-1.47	0.586	0.317	0.33	-0.648	0.567	5.522
ward.6	0.026	118.711	0.722	0.088	-0.944	0.47	0.268	10.005	-1.489	0.488	0.304	0.331	-0.656	0.565	4.601
ward.7	0.025	117.493	0.745	0.088	-0.945	0.47	0.268	9.072	-1.491	0.427	0.293	0.46	-0.657	0.574	4.024
ward.8	0.021	112.814	0.721	0.088	-0.953	0.468	0.261	7.813	-1.489	0.347	0.279	0.445	-0.66	0.603	3.668
ward.9	0.06	108.969	0.686	0.033	-0.794	0.302	0.29	7.071	-0.997	1.742	0.267	0.379	-0.461	0.441	24.133
ward.10	0.057	106.583	0.212	0.033	-0.8	0.302	0.285	6.677	-1.003	1.604	0.258		-0.463	0.445	22.23
ward.11	0.053	106.281	0.203	0.044	-0.809	0.301	0.278	5.78	-1.01	1.467	0.248		-0.467	0.446	20.328
ward.12	0.052	105.544	0.194	0.049	-0.81	0.301	0.277	6.248	-1.011	1.356	0.242		-0.468	0.45	18.786
ward.13	0.051	104.185	0.685	0.049	-0.813	0.301	0.274	5.563	-1.013	1.267	0.235		-0.469	0.457	17.557
ward.14	0.05	103.958	0.266	0.051	-0.814	0.301	0.274	5.107	-1.013	1.179	0.228		-0.469	0.465	16.341

Tabella A.58: R15 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.244	192.535	0.945	0.22	-0.523	0.271	0.641	14.719	-1.204	0.359	0.349	0.473	-0.312	0.528	1.73
ward.3	0.142	276.209	0.875	0.278	-0.665	0.407	0.54	16.226	-1.778	0.247	0.4	0.507	-0.465	0.507	1.188
ward.4	0.07	360.209	0.909	0.286	-0.823	0.442	0.434	19.77	-2.09	0.237	0.401	0.516	-0.573	0.528	0.992
ward.5	0.019	466.93	0.804	0.387	-0.953	0.439	0.347	21.066	-2.291	0.154	0.389	0.585	-0.639	0.572	0.674
ward.6	0.014	494.961	0.397	0.387	-0.965	0.436	0.338	20.388	-2.302	0.134	0.367	0.695	-0.643	0.625	0.54
ward.7	0.01	575.633	0.398	0.387	-0.977	0.432	0.328	23.31	-2.313	0.104	0.349	0.776	-0.646	0.679	0.409
ward.8	0.005	760.074	0.354	0.494	-0.989	0.429	0.317	28.817	-2.325	0.072	0.335	0.843	-0.649	0.739	0.279
ward.9	0.037	1060.995	0.318	0.084	-0.935	0.276	0.255	33.622	-1.855	0.248	0.322	0.779	-0.499	0.657	7.185
ward.10	0.021	1597.184	0.279	0.077	-0.958	0.229	0.205	42.164	-1.755	0.114	0.31	0.765	-0.464	0.712	7.09
ward.11	0.014	1989.475	0.275	0.1	-0.97	0.192	0.169	46.166	-1.64	0.134	0.297	0.744	-0.428	0.714	5.171
ward.12	0.012	2274.247	0.269	0.111	-0.975	0.165	0.146	48.707	-1.545	0.17	0.285	0.732	-0.399	0.735	4.129
ward.13	0.01	2675.795	0.182	0.111	-0.98	0.151	0.13	54.901	-1.49	0.133	0.275	0.737	-0.383	0.758	3.228
ward.14	0.007	3416.328	0.182	0.106	-0.988	0.136	0.111	64.997	-1.434	0.097	0.266	0.746	-0.366	0.787	2.581
ward.15	0.002	4783.717	0.183	0.156	-0.998	0.122	0.089	78.094	-1.376	0.066	0.257	0.75	-0.348	0.818	1.716
ward.16	0.002	4631.968	0.161	0.155	-0.998	0.12	0.087	70.571	-1.369	0.155	0.249	0.72	-0.346	0.807	1.819
ward.17	0.002	4471.249	0.162	0.068	-0.997	0.117	0.087	65.881	-1.346	0.501	0.242	0.668	-0.341	0.78	9.272
ward.18	0.003	4335.804	0.161	0.068	-0.996	0.113	0.086	60.734	-1.328	0.486	0.235	0.636	-0.336	0.763	8.986
ward.19	0.003	4223.037	0.157	0.068	-0.996	0.111	0.085	56.01	-1.316	0.471	0.229	0.615	-0.333	0.751	8.7
ward.20	0.003	4125.478	0.156	0.081	-0.995	0.109	0.084	51.762	-1.304	0.456	0.223	0.587	-0.329	0.735	8.425
ward.21	0.003	4044.358	0.216	0.081	-0.995	0.107	0.083	48.613	-1.292	0.441	0.217	0.558	-0.326	0.712	8.152
ward.22	0.004	3970.206	0.21	0.081	-0.994	0.104	0.082	46.094	-1.269	0.485	0.212	0.543	-0.321	0.705	7.897
ward.23	0.005	3898.828	0.208	0.052	-0.992	0.1	0.081	43.855	-1.244	0.537	0.208	0.512	-0.314	0.678	18.218
ward.24	0.005	3838.451	0.205	0.052	-0.991	0.096	0.08	41.657	-1.22	0.521	0.203	0.496	-0.308	0.665	17.673
ward.25	0.006	3787.631	0.203	0.052	-0.99	0.092	0.08	40.797	-1.194	0.512	0.199	0.469	-0.301	0.635	17.138
ward.26	0.006	3741.359	0.2	0.052	-0.988	0.088	0.079	38.957	-1.168	0.529	0.196	0.455	-0.295	0.624	16.63
ward.27	0.007	3701.245	0.198	0.052	-0.988	0.086	0.078	37.241	-1.152	0.513	0.192	0.439	-0.291	0.601	16.138
ward.28	0.007	3666.319	0.192	0.052	-0.987	0.083	0.077	35.843	-1.133	0.498	0.188	0.434	-0.286	0.598	15.664
ward.29	0.007	3640.009	0.19	0.061	-0.987	0.081	0.076	34.356	-1.117	0.483	0.185	0.43	-0.282	0.587	15.19
ward.30	0.007	3620.824	0.199	0.061	-0.986	0.078	0.075	33.234	-1.097	0.468	0.182	0.424	-0.277	0.585	14.72

Tabella A.59: Transfusion Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.196	409.178	1.116	0.01	-0.619	0.4	0.558	2.522	-0.677	0.405	0.421	0.336	-0.435	0.508	170.324
ward.3	0.159	385.869	0.833	0.013	-0.644	0.407	0.521	18.9	-0.753	0.367	0.412	0.469	-0.453	0.501	129.549
ward.4	0.133	441.667	0.775	0.014	-0.651	0.363	0.498	16.317	-0.655	0.454	0.4	0.445	-0.432	0.492	94.842
ward.5	0.091	444.36	0.864	0.014	-0.75	0.358	0.437	12.341	-0.707	0.373	0.376	0.391	-0.48	0.459	77.75
ward.6	0.08	433.306	0.774	0.018	-0.76	0.314	0.41	10.442	-0.672	0.603	0.352	0.371	-0.454	0.452	67.285

Tabella A.61: S2 Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.214	3166.027	1.002	0.012	-0.557	0.384	0.633	7.957E+10	-93980.362	0.349	0.438	0.416	-0.391	0.504	429.717
ward.3	0.127	4174.647	0.838	0.014	-0.722	0.402	0.512	7.7606E+10	-117951.84	0.219	0.458	0.459	-0.493	0.524	262.808
ward.4	0.087	4548.23	0.701	0.018	-0.805	0.363	0.44	1.05E+11	-122034.78	0.245	0.427	0.458	-0.511	0.524	188.127
ward.5	0.079	4572.789	0.621	0.018	-0.822	0.319	0.408	9.0726E+10	-117177.28	0.32	0.396	0.443	-0.486	0.491	150.566
ward.6	0.058	5011.464	0.858	0.02	-0.872	0.268	0.35	7.9271E+10	-113681.9	0.272	0.373	0.436	-0.467	0.505	162.522
ward.7	0.047	5232.294	0.502	0.021	-0.898	0.246	0.319	7.2695E+10	-112181.53	0.244	0.351	0.461	-0.457	0.528	134.198
ward.8	0.039	5711.304	0.454	0.021	-0.92	0.225	0.291	8.6032E+10	-110016.15	0.197	0.333	0.485	-0.445	0.541	108.56
ward.9	0.034	6139.178	0.306	0.022	-0.932	0.213	0.274	1.02E+11	-108755.77	0.168	0.317	0.531	-0.438	0.573	90.216
ward.10	0.031	6655.542	0.314	0.023	-0.941	0.202	0.259	1.06E+11	-107296.3	0.192	0.304	0.575	-0.429	0.613	75.206
ward.11	0.026	7427.695	0.319	0.021	-0.957	0.178	0.229	1.08E+11	-103284.46	0.191	0.292	0.574	-0.408	0.618	86.362
ward.12	0.022	8359.6	0.321	0.022	-0.967	0.165	0.209	1.15E+11	-101037.91	0.156	0.281	0.592	-0.395	0.645	70.599
ward.13	0.018	9321.256	0.309	0.022	-0.974	0.152	0.191	1.23E+11	-98561.79	0.156	0.271	0.599	-0.382	0.668	58.564
ward.14	0.015	10609.717	0.307	0.022	-0.98	0.139	0.172	1.49E+11	-95905.383	0.14	0.263	0.616	-0.367	0.697	47.871
ward.15	0.01	12378.993	0.292	0.024	-0.99	0.124	0.147	1.63E+11	-92602.143	0.134	0.255	0.61	-0.35	0.713	45.848
ward.16	0.01	11979.491	0.296	0.016	-0.989	0.121	0.146	1.50E+11	-91571.553	0.347	0.247	0.575	-0.346	0.7	102.468
ward.17	0.011	11646.316	0.302	0.017	-0.989	0.119	0.143	1.36E+11	-90716.493	0.335	0.239	0.54	-0.342	0.685	98.888
ward.18	0.012	11351.27	0.312	0.017	-0.988	0.114	0.14	1.25E+11	-88917.312	0.524	0.233	0.507	-0.334	0.667	95.557
ward.19	0.012	11081.696	0.316	0.017	-0.988	0.111	0.138	1.15E+11	-88141.696	0.508	0.227	0.492	-0.331	0.657	92.501
ward.20	0.013	10815.028	0.31	0.017	-0.988	0.108	0.136	1.07E+11	-86792.305	0.493	0.221	0.477	-0.325	0.652	89.84
ward.21	0.013	10577.222	0.628	0.017	-0.987	0.105	0.134	9.8999E+10	-85740.537	0.479	0.216	0.455	-0.321	0.635	87.309
ward.22	0.013	10370.174	0.539	0.012	-0.987	0.103	0.132	9.2193E+10	-84977.191	0.466	0.211	0.447	-0.318	0.634	155.535
ward.23	0.013	10198.487	0.557	0.012	-0.987	0.1	0.13	8.6337E+10	-84048.733	0.452	0.206	0.434	-0.314	0.622	151.035
ward.24	0.013	10024.372	0.573	0.012	-0.987	0.099	0.128	8.1219E+10	-83391.145	0.44	0.202	0.432	-0.311	0.619	147.034
ward.25	0.013	9875.874	0.57	0.015	-0.988	0.098	0.127	7.6027E+10	-83138.265	0.428	0.198	0.429	-0.31	0.615	143.081
ward.26	0.013	9753.09	0.568	0.015	-0.988	0.096	0.125	7.1947E+10	-82325.221	0.417	0.194	0.419	-0.306	0.606	139.139
ward.27	0.013	9637.154	0.559	0.015	-0.987	0.093	0.123	6.823E+10	-81243.31	0.46	0.191	0.408	-0.302	0.597	135.443
ward.28	0.013	9531.158	0.557	0.015	-0.987	0.091	0.121	6.4941E+10	-80497.668	0.448	0.187	0.405	-0.299	0.589	131.919
ward.29	0.013	9425.013	0.57	0.015	-0.987	0.09	0.12	6.1877E+10	-80242.511	0.437	0.184	0.405	-0.298	0.587	128.675
ward.30	0.014	9321.169	0.582	0.013	-0.987	0.087	0.119	5.9408E+10	-78861.462	0.539	0.181	0.39	-0.292	0.567	159.138

Tabella A.62: Wine Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.184	65.361	1.412	0.216	-0.599	0.393	0.743	5.77	-0.73	0.581	0.368	0.29	-0.42	0.405	1.61
ward.3	0.126	67.647	1.259	0.229	-0.742	0.386	0.663	4.934	-0.874	0.583	0.381	0.276	-0.494	0.405	1.864
ward.4	0.13	51.464	1.893	0.211	-0.745	0.35	0.653	2.946	-0.84	1.263	0.343	0.199	-0.471	0.351	2.054
ward.5	0.136	43.679	1.503	0.191	-0.741	0.31	0.647	2.318	-0.795	1.186	0.317	0.168	-0.442	0.306	2.351
ward.6	0.115	39.129	1.384	0.191	-0.788	0.296	0.622	1.965	-0.816	1.115	0.298	0.158	-0.453	0.298	2.211

Tabella A.63: Wineqred Ward

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
ward.2	0.16	1058.705	1.115	0.012	-0.642	0.406	0.538	1315.814	-15.944	0.336	0.22	0.35	-0.451	0.492	90.933
ward.3	0.145	1011.505	0.801	0.012	-0.69	0.381	0.499	1207.363	-14.835	0.436	0.226	0.391	-0.463	0.516	66.687
ward.4	0.072	1092.085	0.981	0.012	-0.845	0.378	0.393	1585.635	-17.178	0.324	0.22	0.386	-0.541	0.542	49.513
ward.5	0.052	1114.976	0.784	0.018	-0.883	0.374	0.362	2243.125	-17.744	0.26	0.205	0.396	-0.556	0.535	39.815
ward.6	0.044	1175.711	0.57	0.018	-0.896	0.375	0.352	2034.817	-18.047	0.211	0.225	0.42	-0.563	0.546	32.241
ward.7	0.038	1154.814	0.561	0.018	-0.913	0.368	0.339	1682.96	-18.074	0.269	0.21	0.398	-0.566	0.513	28.252
ward.8	0.041	1154.484	0.549	0.013	-0.904	0.336	0.331	1512.716	-17.127	0.442	0.2	0.37	-0.537	0.485	46.514
ward.9	0.043	1185.736	0.553	0.012	-0.896	0.24	0.298	1423.241	-14.385	0.678	0.191	0.331	-0.451	0.433	50.51
ward.10	0.041	1226.984	0.133	0.022	-0.898	0.24	0.295	5820.131	-14.441	0.594	0.183	0.396	-0.452	0.435	44.26
ward.11	0.042	1219.873	0.135	0.022	-0.893	0.204	0.285	5552.15	-13.278	0.592	0.176	0.389	-0.415	0.431	40.527
ward.12	0.04	1200.11	0.135	0.024	-0.9	0.202	0.279	4929.946	-13.282	0.552	0.169	0.377	-0.415	0.421	37.759

Tabella A.64: fbis SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.651	26.933	3.94	0.004	0.067	0.22	1.183	108.437	5.798	17.321	0.058	0.049	0.046	0.065	92.555
skmeans.3	0.398	36.566	3.81	0.003	-0.03	0.232	1.016	103.499	0.538	25.867	0.075	-0.074	-0.02	0.086	220.282
skmeans.4	0.382	28.063	3.709	0.005	-0.003	0.208	1.067	73.708	2.176	32.368	0.071	-0.07	-0.002	0.071	90.47
skmeans.5	0.266	36.651	3.184	0.003	-0.118	0.2	0.876	139.574	-3.924	34.924	0.078	-0.088	-0.071	0.074	214.063
skmeans.6	0.256	30.444	3.044	0.005	-0.095	0.169	0.927	96.713	-2.11	32.938	0.075	-0.075	-0.053	0.064	88.109
skmeans.7	0.248	25.377	4.806	0.003	-0.084	0.148	0.966	61.062	-0.913	29.366	0.072	-0.071	-0.044	0.077	293.685
skmeans.8	0.23	22.455	4.605	0.002	-0.084	0.131	0.966	54.949	-0.855	28.429	0.069	-0.084	-0.041	0.075	293.124
skmeans.9	0.209	22.46	4.689	0.003	-0.091	0.114	0.959	45.586	-0.965	28.68	0.07	-0.13	-0.042	0.063	290.614
skmeans.10	0.199	19.207	4.267	0.002	-0.102	0.107	0.95	35.901	-1.13	38.572	0.067	-0.093	-0.045	0.071	291.36
skmeans.11	0.188	19.098	3.891	0.003	-0.117	0.099	0.94	33.576	-1.298	35.418	0.067	-0.16	-0.049	0.055	289.355
skmeans.12	0.173	19.923	4.343	0.002	-0.152	0.105	0.874	34.383	-2.804	28.272	0.067	-0.086	-0.065	0.078	286.294
skmeans.13	0.159	19.734	3.49	0.003	-0.169	0.095	0.858	32.74	-2.982	27.556	0.067	-0.099	-0.068	0.072	284.403
skmeans.14	0.166	18.518	4.115	0.002	-0.132	0.092	0.885	27.786	-2.406	29.012	0.063	-0.168	-0.053	0.069	283.977
skmeans.15	0.15	18.824	3.609	0.003	-0.163	0.079	0.874	30.282	-2.41	70.497	0.064	-0.174	-0.06	0.044	281.58
skmeans.16	0.158	15.709	3.554	0.003	-0.121	0.073	0.924	20.944	-1.413	27.246	0.061	-0.188	-0.044	0.045	310.359
skmeans.17	0.156	15.708	2.933	0.003	-0.121	0.069	0.931	18.888	-1.252	84.296	0.059	-0.167	-0.043	0.035	282.832
skmeans.18	0.158	13.386	4.486	0.003	-0.104	0.065	0.954	11.69	-0.805	68.798	0.057	-0.177	-0.036	0.036	285.335
skmeans.19	0.152	14.557	3.45	0.002	-0.12	0.063	0.944	16.585	-0.965	74.767	0.058	-0.184	-0.04	0.037	397.682
skmeans.20	0.139	14.819	2.527	0.002	-0.158	0.066	0.877	17.375	-2.137	61.413	0.058	-0.166	-0.053	0.044	394.814
skmeans.21	0.137	14.291	3.764	0.002	-0.15	0.06	0.895	15.384	-1.744	71.229	0.058	-0.194	-0.048	0.037	394.182
skmeans.22	0.128	14.88	2.694	0.003	-0.145	0.056	0.868	24.232	-2.148	61.234	0.058	-0.16	-0.046	0.039	315.996
skmeans.23	0.132	13.636	2.735	0.003	-0.145	0.055	0.894	16.02	-1.688	98.648	0.056	-0.177	-0.045	0.037	277.744
skmeans.24	0.13	13.064	3.74	0.003	-0.135	0.05	0.911	14.855	-1.364	101.804	0.056	-0.216	-0.04	0.039	277.683
skmeans.25	0.14	12.607	3.702	0.003	-0.107	0.05	0.954	12.119	-0.703	118.459	0.055	-0.208	-0.032	0.035	317.094
skmeans.26	0.139	11.399	3.193	0.003	-0.108	0.048	0.963	9.367	-0.554	66.638	0.054	-0.197	-0.032	0.031	279.239
skmeans.27	0.119	12.282	3.552	0.003	-0.148	0.048	0.866	11.344	-2.006	90.888	0.053	-0.172	-0.043	0.038	275.746
skmeans.28	0.127	12.207	3.165	0.003	-0.116	0.046	0.914	10.821	-1.273	113.708	0.053	-0.184	-0.033	0.033	274.71
skmeans.29	0.116	11.423	3.456	0.002	-0.192	0.053	0.83	9.799	-2.641	97.445	0.053	-0.181	-0.058	0.046	389.178
skmeans.30	0.113	12.57	2.741	0.002	-0.188	0.048	0.843	21.978	-2.325	95.612	0.054	-0.171	-0.054	0.043	271.251
skmeans.31	0.127	11.046	3.246	0.005	-0.128	0.044	0.929	9.076	-1.023	97.896	0.052	-0.22	-0.036	0.038	101.351
skmeans.32	0.116	10.567	2.489	0.004	-0.172	0.045	0.875	8.217	-1.785	100.796	0.051	-0.192	-0.047	0.043	194.015
skmeans.33	0.102	11.501	2.986	0.003	-0.191	0.042	0.818	14.56	-2.498	119.288	0.053	-0.223	-0.051	0.04	270.869
skmeans.34	0.119	10.134	2.407	0.003	-0.133	0.039	0.917	7.527	-1.119	88.18	0.05	-0.194	-0.035	0.036	274.149

Tabella A.65: la1 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.442	57.367	6.025	0.02	-0.184	0.296	0.91	16.297	-1.76	13.952		0.135	-0.13	0.105	51.73
skmeans.3	0.354	56.473	5.729	0.02	-0.312	0.31	0.811	16.77	-3.68	12.979		-0.009	-0.214	0.106	50.863
skmeans.4	0.287	43.271	7.766	0.027	-0.395	0.284	0.74	11.495	-4.723	15.311		-0.08	-0.252	0.09	28.676
skmeans.5	0.258	37.498	4.674	0.027	-0.442	0.278	0.704	8.561	-5.244	14.477		-0.119	-0.275	0.092	28.503
skmeans.6	0.319	30.145	6.537	0.028	-0.248	0.183	0.835	5.683	-2.418	96.53		-0.189	-0.134	0.024	28.496
skmeans.7	0.337	26.398	7.374	0.027	-0.205	0.171	0.871	4.685	-1.847	98.087		-0.162	-0.11	0.028	28.431
skmeans.8	0.268	25.338	3.796	0.028	-0.358	0.189	0.755	3.972	-3.547	83.505		-0.169	-0.189	0.032	28.27
skmeans.9	0.271	24.726	5.272	0.028	-0.329	0.161	0.777	3.738	-2.985	83.112		-0.186	-0.162	0.03	28.099
skmeans.10	0.26	22.771	6.626	0.02	-0.337	0.147	0.768	2.82	-2.948	148.14		-0.186	-0.158	0.019	60.085
skmeans.11	0.281	21.167	5.845	0.028	-0.252	0.116	0.827	2.58	-1.994	125.501		-0.184	-0.109	0.025	27.984
skmeans.12	0.244	19.72	5.353	0.029	-0.367	0.141	0.746	2.191	-3.132	83.548		-0.175	-0.166	0.034	27.94

Tabella A.66: la2 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.483	52.433	6.062	0.027	-0.089	0.272	0.974	14.903	-0.495	14.559		0.132	-0.063	0.095	36.515
skmeans.3	0.392	53.871	4.83	0.028	-0.226	0.287	0.871	15.346	-2.415	12.295		-0.017	-0.154	0.101	35.88
skmeans.4	0.295	43.406	5.436	0.039	-0.379	0.279	0.753	11.461	-4.399	13.406		-0.082	-0.241	0.094	21.011
skmeans.5	0.282	36.435	4.846	0.039	-0.382	0.245	0.747	7.875	-4.164	20.362		-0.127	-0.227	0.068	20.91
skmeans.6	0.268	32.099	5.625	0.039	-0.391	0.222	0.734	6.052	-4.144	19.138		-0.154	-0.221	0.066	20.814
skmeans.7	0.268	27.795	4.477	0.039	-0.388	0.215	0.736	4.72	-4.037	31.63		-0.176	-0.216	0.052	20.773
skmeans.8	0.312	25.233	4.235	0.036	-0.257	0.161	0.825	3.94	-2.374	83.499		-0.164	-0.13	0.03	24.475
skmeans.9	0.301	23.295	5.292	0.034	-0.265	0.145	0.817	3.15	-2.347	80.831		-0.177	-0.127	0.03	24.401
skmeans.10	0.278	22.48	5.342	0.021	-0.313	0.136	0.787	2.751	-2.603	84.661		-0.216	-0.143	0.022	66.774
skmeans.11	0.253	20.802	6.677	0.022	-0.38	0.157	0.736	2.235	-3.399	33.041		-0.212	-0.181	0.046	66.657
skmeans.12	0.263	20.184	5.775	0.037	-0.337	0.128	0.767	2.145	-2.732	76.349		-0.206	-0.147	0.033	24.135

Tabella A.67: k1a SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.4	49.575	5.511	0.14	-0.317	0.292	0.899	15.898	-1.548	7.819		0.037	-0.211	0.081	2.235
skmeans.3	0.319	42.222	5.807	0.098	-0.318	0.323	0.869	10.297	-2.083	10.293		-0.007	-0.223	0.063	4.488
skmeans.4	0.299	38.631	5.139	0.094	-0.333	0.295	0.861	7.639	-2.07	9.523		-0.045	-0.221	0.06	4.887
skmeans.5	0.289	35.122	4.557	0.112	-0.337	0.283	0.856	6.914	-2.103	9.368		-0.055	-0.22	0.063	4.838
skmeans.6	0.303	30.359	4.484	0.112	-0.294	0.247	0.871	4.783	-1.77	11.232		-0.06	-0.182	0.054	4.816
skmeans.7	0.309	26.2	5.589	0.112	-0.261	0.201	0.882	2.885	-1.458	13.323		-0.075	-0.147	0.049	4.806
skmeans.8	0.26	28.16	4.766	0.072	-0.371	0.237	0.837	5.602	-2.13	12.902		-0.07	-0.218	0.05	12.4
skmeans.9	0.287	26.342	5.222	0.092	-0.289	0.2	0.864	3.916	-1.668	11.148		-0.072	-0.161	0.055	7.605
skmeans.10	0.288	24.895	2.23	0.09	-0.283	0.187	0.866	4.955	-1.584	13.094		-0.041	-0.153	0.055	5.1
skmeans.11	0.242	24.813	5.559	0.068	-0.384	0.169	0.827	7.803	-1.874	17.625		-0.095	-0.19	0.043	14.052
skmeans.12	0.262	22.089	4.334	0.072	-0.339	0.167	0.846	2.584	-1.687	11.62		-0.081	-0.17	0.055	12.177
skmeans.13	0.257	22.747	3.583	0.104	-0.367	0.166	0.842	7.182	-1.702	12.234		-0.02	-0.181	0.056	5.365
skmeans.14	0.261	22.252	5.535	0.076	-0.332	0.142	0.85	6.86	-1.51	16.028		-0.064	-0.154	0.049	9.219
skmeans.15	0.233	19.837	4.34	0.061	-0.41	0.156	0.82	2.75	-1.852	13.349		-0.058	-0.193	0.054	14.338
skmeans.16	0.267	19.712	6.252	0.064	-0.326	0.14	0.857	4.784	-1.436	12.31		-0.028	-0.15	0.058	13.794
skmeans.17	0.223	20.253	4.346	0.098	-0.426	0.147	0.812	3.549	-1.872	12.538		-0.061	-0.193	0.056	5.671
skmeans.18	0.25	21.711	6.239	0.067	-0.34	0.124	0.844	14.336	-1.459	21.945		-0.022	-0.146	0.048	13.416
skmeans.19	0.203	19.262	3.929	0.064	-0.462	0.131	0.793	3.558	-1.915	18.368		-0.058	-0.196	0.053	12.732
skmeans.20	0.221	22.136	4.466	0.066	-0.4	0.123	0.813	13.008	-1.707	14.357		-0.016	-0.168	0.06	14.04
skmeans.21	0.236	16.848	5.25	0.059	-0.352	0.107	0.834	1.232	-1.434	17.049		-0.076	-0.14	0.052	18.102
skmeans.22	0.199	18.279	5.029	0.067	-0.491	0.137	0.788	2.745	-1.989	14.625		-0.04	-0.21	0.061	14.229
skmeans.23	0.193	15.888	4.479	0.054	-0.48	0.111	0.785	1.093	-1.812	14.684		-0.081	-0.186	0.058	21.617
skmeans.24	0.2	17.538	6.367	0.068	-0.452	0.103	0.796	2.85	-1.674	17.669		-0.074	-0.17	0.057	13.243
skmeans.25	0.197	19.025	4.514	0.11	-0.46	0.111	0.79	8.422	-1.782	23.45		-0.046	-0.179	0.058	5.007
skmeans.26	0.218	15.931	3.29	0.081	-0.35	0.075	0.828	2.077	-1.237	20.093		-0.09	-0.116	0.055	9.648
skmeans.27	0.217	16.454	3.35	0.07	-0.368	0.079	0.825	2.039	-1.29	37.192		-0.084	-0.125	0.056	7.776
skmeans.28	0.199	15.298	5.815	0.082	-0.44	0.097	0.796	1.525	-1.62	14.551		-0.068	-0.161	0.065	7.818
skmeans.29	0.191	16.481	1.528	0.076	-0.474	0.1	0.786	6.094	-1.717	14.851		-0.032	-0.175	0.064	10.369
skmeans.30	0.219	16.773	2.82	0.068	-0.351	0.07	0.832	5.922	-1.162	44.944		-0.067	-0.113	0.056	12.844
skmeans.31	0.206	16.695	3.657	0.066	-0.393	0.077	0.811	5.588	-1.36	32.244		-0.038	-0.131	0.062	13.629
skmeans.32	0.189	14.308	3.997	0.066	-0.455	0.083	0.789	1.413	-1.548	26.363		-0.068	-0.154	0.06	13.912
skmeans.33	0.217	14.358	3.764	0.054	-0.352	0.073	0.828	1.46	-1.224	39.006		-0.082	-0.116	0.059	12.967
skmeans.34	0.228	15.788	3.784	0.074	-0.32	0.073	0.841	4.666	-1.138	37.747		-0.063	-0.106	0.062	10.681
skmeans.35	0.206	14.964	3.26	0.076	-0.405	0.068	0.817	2.577	-1.232	40.413		-0.073	-0.126	0.059	10.19
skmeans.36	0.17	15.388	4.832	0.067	-0.526	0.093	0.762	4.181	-1.807	12.739		-0.038	-0.184	0.074	13.443
skmeans.37	0.207	12.895	3.851	0.069	-0.396	0.068	0.82	1.103	-1.221	41.796		-0.079	-0.124	0.059	12.941
skmeans.38	0.176	13.302	1.902	0.071	-0.493	0.082	0.773	1.337	-1.626	15.649		-0.073	-0.163	0.067	12.057
skmeans.39	0.169	14.069	4.522	0.07	-0.535	0.093	0.761	3.572	-1.809	14.854		-0.035	-0.186	0.074	11.875
skmeans.40	0.192	14.126	3.218	0.069	-0.426	0.065	0.801	3.41	-1.297	36		-0.064	-0.129	0.065	12.544

Tabella A.68: k1b SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.378	49.879	5.575	0.15	-0.331	0.297	0.883	15.83	-1.816	7.853		0.038	-0.221	0.081	2.235
skmeans.3	0.315	42.54	5.128	0.098	-0.327	0.326	0.865	10.594	-2.146	9.888		-0.008	-0.23	0.065	4.487
skmeans.4	0.302	38.433	5.673	0.094	-0.327	0.291	0.864	7.618	-2.022	9.569		-0.044	-0.217	0.06	4.888
skmeans.5	0.33	32.556	4.254	0.086	-0.261	0.267	0.888	5.487	-1.614	9.866		-0.031	-0.17	0.058	7.855
skmeans.6	0.28	30.758	4.501	0.108	-0.343	0.255	0.851	4.783	-2.038	9.366		-0.065	-0.212	0.06	4.812
skmeans.7	0.285	30.154	4.732	0.095	-0.316	0.223	0.859	6.488	-1.813	19.695		-0.071	-0.184	0.044	7.215
skmeans.8	0.278	25.076	4.563	0.087	-0.346	0.223	0.853	2.915	-1.862	11.33		-0.059	-0.199	0.052	7.712
skmeans.9	0.288	25.018	6.964	0.069	-0.288	0.163	0.871	3.973	-1.42	26.459		-0.088	-0.145	0.034	13.477
skmeans.10	0.324	22.655	4.533	0.069	-0.2	0.144	0.91	3.278	-0.955	24.094		-0.065	-0.098	0.037	12.366
skmeans.11	0.32	22.728	4.784	0.065	-0.181	0.121	0.914	2.806	-0.838	21.507		-0.064	-0.082	0.034	14.167
skmeans.12	0.295	22.837	4.496	0.096	-0.232	0.15	0.879	2.914	-1.296	18.798		-0.09	-0.115	0.045	6.909

Tabella A.69: re0 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.489	56.771	4.213	0.029	-0.03	0.256	0.971	10.176	-0.301	6.16	0.037	0.04	-0.021	0.106	48.748
skmeans.3	0.43	51.733	3.919	0.033	-0.053	0.235	0.958	6.315	-0.403	5.517	0.051	0.008	-0.035	0.098	47.328
skmeans.4	0.398	47.512	3.898	0.021	-0.07	0.201	0.943	4.868	-0.51	4.722	0.05	0.025	-0.043	0.101	115.501
skmeans.5	0.394	43.863	2.475	0.021	-0.065	0.183	0.949	4.084	-0.431	4.027	0.051	-0.024	-0.038	0.122	113.223
skmeans.6	0.374	40.253	3.018	0.021	-0.084	0.157	0.934	3.157	-0.514	5.417	0.051	-0.038	-0.045	0.118	111.496
skmeans.7	0.377	37.64	3.87	0.021	-0.063	0.14	0.947	2.397	-0.393	5.216	0.049	-0.061	-0.032	0.07	109.896
skmeans.8	0.379	33.159	3.002	0.021	-0.053	0.132	0.954	1.939	-0.331	5.242	0.047	-0.091	-0.026	0.075	109.488
skmeans.9	0.37	30.659	3.385	0.03	-0.066	0.121	0.945	1.583	-0.376	6.502	0.047	-0.111	-0.031	0.068	54.325
skmeans.10	0.342	30.052	4.314	0.021	-0.12	0.105	0.907	1.917	-0.584	5.777	0.048	-0.079	-0.052	0.062	107.089
skmeans.11	0.34	27.198	4.816	0.033	-0.126	0.103	0.904	1.318	-0.591	6.239	0.048	-0.125	-0.054	0.063	42.794
skmeans.12	0.336	26.25	2.334	0.015	-0.128	0.092	0.901	1.397	-0.575	6.555	0.048	-0.128	-0.052	0.06	211.936
skmeans.13	0.349	23.807	3.711	0.033	-0.095	0.099	0.923	1.085	-0.472	5.929	0.045	-0.115	-0.041	0.072	42.455
skmeans.14	0.33	23.563	3.91	0.015	-0.14	0.085	0.893	1.025	-0.593	6.547	0.047	-0.134	-0.054	0.058	209.816
skmeans.15	0.328	22.506	4.034	0.016	-0.142	0.08	0.891	0.896	-0.588	7.095	0.046	-0.108	-0.053	0.065	208.773
skmeans.16	0.334	21.382	2.109	0.021	-0.117	0.074	0.906	0.849	-0.494	6.998	0.044	-0.127	-0.043	0.067	104.048
skmeans.17	0.314	20.993	3.927	0.032	-0.182	0.08	0.865	0.736	-0.714	7.752	0.046	-0.141	-0.067	0.067	51.585
skmeans.18	0.32	20.317	2.874	0.021	-0.151	0.067	0.883	0.67	-0.574	7.487	0.045	-0.14	-0.051	0.067	102.623
skmeans.19	0.312	18.69	3.85	0.031	-0.174	0.069	0.866	0.738	-0.664	6.061	0.045	-0.144	-0.06	0.062	51.557
skmeans.20	0.317	18.832	3.056	0.023	-0.158	0.063	0.879	0.621	-0.572	6.106	0.044	-0.104	-0.052	0.074	101.905
skmeans.21	0.318	18.167	3.733	0.023	-0.147	0.058	0.883	0.513	-0.534	8.621	0.044	-0.106	-0.047	0.073	101.586
skmeans.22	0.317	17.033	3.908	0.015	-0.162	0.061	0.879	0.457	-0.567	6.553	0.043	-0.133	-0.053	0.058	203.771
skmeans.23	0.31	17.42	4.018	0.023	-0.162	0.054	0.868	0.552	-0.579	9.048	0.043	-0.101	-0.05	0.068	100.476
skmeans.24	0.299	16.505	3.334	0.023	-0.206	0.061	0.843	0.397	-0.719	9.09	0.043	-0.182	-0.065	0.047	100.657
skmeans.25	0.311	16.12	2.203	0.021	-0.167	0.052	0.871	0.425	-0.553	8.305	0.043	-0.126	-0.05	0.063	100.252
skmeans.26	0.306	16.242	3.531	0.021	-0.189	0.055	0.859	0.49	-0.618	13.35	0.045	-0.199	-0.058	0.048	99.218

Tabella A.70: tr23 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.784	2.11	2.006	0.008	-0.492	0.265	1.22	7502.175	26.649	10.994	0.057	0.156	-0.293	0.08	199.793
skmeans.3	0.222	12.174	1.489	0.003	-0.603	0.396	0.395	27924.192	-138.634	21.163	0.147	0.08	-0.424	0.528	1131.851
skmeans.4	0.224	8.522	2.291	0.002	-0.392	0.279	0.712	17194.954	-47.015	1280.436	0.133	-0.074	-0.248	0.174	3428.337
skmeans.5	0.233	6.362	1.758	0.002	-0.231	0.213	0.93	11048.966	-9.909	2244.006	0.119	-0.087	-0.136	0.128	4104.179
skmeans.6	0.115	11.951	2.232	0.002	-0.445	0.214	0.525	35092.962	-67.083	907.872	0.152	-0.097	-0.242	0.188	2970.207
skmeans.7	0.111	10.425	2.996	0.002	-0.435	0.203	0.531	28389.448	-64.348	940.675	0.144	-0.11	-0.231	0.193	2934.755
skmeans.8	0.12	8.455	2.938	0.002	-0.353	0.173	0.636	19737.315	-46.146	1506.793	0.133	-0.138	-0.178	0.127	3555.395
skmeans.9	0.129	7.365	1.298	0.002	-0.269	0.128	0.837	15697.283	-17.654	3237.983	0.126	-0.148	-0.121	0.153	3554.946
skmeans.10	0.13	6.515	1.307	0.002	-0.284	0.126	0.859	12748.403	-15.066	3382.584	0.121	-0.143	-0.126	0.158	3554.699
skmeans.11	0.125	6.09	1.225	0.002	-0.283	0.122	0.845	11396.248	-16.292	3424.335	0.117	-0.153	-0.123	0.156	3518.734
skmeans.12	0.127	5.548	2.072	0.002	-0.276	0.117	0.88	9714.624	-12.346	3201.691	0.113	-0.167	-0.118	0.156	3512.542

Tabella A.71: tr41 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.45	9.524	4.629	0.01	-0.072	0.268	0.894	68.154	-3.201	22.426	0.049	0.031	-0.051	0.115	29.479
skmeans.3	0.389	6.904	4.384	0.006	-0.094	0.25	0.968	40.316	-0.884	26.364	0.056	-0.039	-0.064	0.08	81.981
skmeans.4	0.379	5.413	4.055	0.006	-0.079	0.219	1.055	26.159	1.401	49.741	0.055	-0.118	-0.05	0.037	81.756
skmeans.5	0.338	5.225	3.632	0.005	-0.096	0.197	1.05	22.387	1.197	34.403	0.054	-0.117	-0.058	0.046	104.157
skmeans.6	0.272	5.236	4.216	0.005	-0.079	0.158	1.013	19.641	0.28	31.637	0.054	-0.102	-0.043	0.063	103.542
skmeans.7	0.16	9.651	3.08	0.004	-0.248	0.189	0.705	93.209	-6.953	51.827	0.064	-0.162	-0.137	0.085	158.337
skmeans.8	0.177	6.804	3.471	0.005	-0.176	0.152	0.805	58.482	-4.136	50.231	0.058	-0.145	-0.09	0.077	101.115
skmeans.9	0.158	6.315	3.168	0.006	-0.231	0.136	0.806	43.755	-3.792	55.783	0.058	-0.165	-0.109	0.068	83.263
skmeans.10	0.195	5.078	3.624	0.005	-0.097	0.122	0.922	37.429	-1.511	80.668	0.053	-0.158	-0.046	0.059	101.316
skmeans.11	0.174	7.417	1.048	0.005	-0.149	0.137	0.827	203.502	-3.502	72.331	0.053	-0.185	-0.073	0.068	98.246
skmeans.12	0.12	6.06	3.944	0.005	-0.257	0.109	0.755	31.803	-4.223	50.619	0.057	-0.196	-0.107	0.069	99.029
skmeans.13	0.123	7.805	3.291	0.006	-0.216	0.105	0.766	157.014	-4.033	57.638	0.055	-0.114	-0.09	0.074	79.495
skmeans.14	0.12	5.818	3.678	0.006	-0.228	0.105	0.757	42.308	-4.158	41.054	0.055	-0.165	-0.094	0.075	81.011
skmeans.15	0.121	7.122	3.611	0.003	-0.217	0.108	0.751	79.188	-4.34	46.776	0.06	-0.122	-0.091	0.086	227.061
skmeans.16	0.105	6.556	3.796	0.006	-0.298	0.092	0.754	107.867	-3.814	53.359	0.054	-0.129	-0.112	0.081	74.748
skmeans.17	0.142	3.521	3.128	0.006	-0.186	0.08	0.935	9.959	-0.968	85.112	0.052	-0.152	-0.068	0.077	82.692
skmeans.18	0.111	5.812	3.216	0.005	-0.204	0.082	0.79	84.978	-3.174	75.147	0.052	-0.116	-0.075	0.078	95.66
skmeans.19	0.092	6.822	4.21	0.005	-0.29	0.086	0.711	86.215	-4.366	72.154	0.055	-0.14	-0.106	0.083	93.312
skmeans.20	0.095	6.013	3.462	0.006	-0.25	0.072	0.77	76.039	-3.188	80.322	0.052	-0.154	-0.085	0.081	73.49

Tabella A.72: tr45 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.601	3.276	4.045	0.002	-0.073	0.261	1.093	327.72	5.306	43.607	0.044	0.019	-0.051	0.077	637.812
skmeans.3	0.323	8.012	3.778	0.002	-0.087	0.244	0.877	673.57	-7.391	148.138	0.068	-0.163	-0.059	0.084	626.242
skmeans.4	0.169	12.246	3.661	0.002	-0.252	0.256	0.564	1891.635	-27.339	208.437	0.087	-0.185	-0.161	0.079	608.275
skmeans.5	0.276	5.021	2.041	0.002	-0.153	0.238	0.851	504.91	-8.581	231.133	0.062	-0.199	-0.099	0.079	622.595
skmeans.6	0.219	5.144	3.078	0.002	-0.208	0.199	0.898	475.545	-5.117	238.884	0.064	-0.197	-0.119	0.076	617.625
skmeans.7	0.117	10.091	3.406	0.002	-0.288	0.184	0.599	1336.881	-19.798	284.166	0.08	-0.154	-0.154	0.075	588.665
skmeans.8	0.111	8.91	3.169	0.002	-0.319	0.169	0.627	1042.592	-17.162	302.966	0.076	-0.125	-0.161	0.094	587.155
skmeans.9	0.109	7.927	3.561	0.002	-0.331	0.149	0.678	851.167	-13.683	291.888	0.073	-0.142	-0.157	0.098	1057.559
skmeans.10	0.102	6.104	2.805	0.002	-0.345	0.14	0.673	433.246	-13.351	247.697	0.069	-0.176	-0.158	0.099	592.947
skmeans.11	0.101	6.018	3.011	0.002	-0.319	0.125	0.706	585.048	-11.353	228.512	0.068	-0.21	-0.139	0.07	1061.923
skmeans.12	0.107	6.008	3.286	0.002	-0.313	0.12	0.759	522.747	-9.079	253.503	0.067	-0.173	-0.134	0.086	1053.358
skmeans.13	0.095	5.327	3.307	0.002	-0.334	0.119	0.7	451.056	-11.189	218.144	0.066	-0.203	-0.141	0.072	1056.308
skmeans.14	0.095	4.992	3.839	0.002	-0.245	0.093	0.765	380.5	-7.987	233.519	0.064	-0.22	-0.095	0.07	584.719
skmeans.15	0.075	5.883	2.583	0.002	-0.349	0.1	0.64	410.48	-12.248	392.091	0.068	-0.18	-0.134	0.092	1030.33
skmeans.16	0.069	9.07	2.51	0.002	-0.308	0.098	0.595	1895.447	-13.892	281.131	0.073	-0.198	-0.119	0.08	533.221
skmeans.17	0.067	8.023	3.004	0.002	-0.282	0.083	0.629	1460.802	-11.743	494.308	0.069	-0.186	-0.101	0.083	538.199
skmeans.18	0.058	5.842	2.319	0.002	-0.345	0.083	0.574	429.969	-13.247	336.991	0.068	-0.193	-0.121	0.089	1007.189
skmeans.19	0.068	6.081	1.957	0.002	-0.315	0.084	0.636	1237.815	-11.483	349.225	0.064	-0.225	-0.113	0.088	993.908
skmeans.20	0.097	3.356	3.346	0.002	-0.218	0.069	0.907	181.649	-2.711	573.66	0.056	-0.217	-0.074	0.069	585.16

Tabella A.73: wap SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.404	35.584	5.362	0.141	-0.319	0.297	0.896	16.046	-1.562	7.479		0.039	-0.214	0.085	2.184
skmeans.3	0.296	31.62	5.185	0.119	-0.366	0.339	0.846	11.986	-2.445	7.545		-0.012	-0.258	0.077	3.017
skmeans.4	0.29	28.785	4.972	0.094	-0.342	0.298	0.852	7.997	-2.172	8.328		-0.047	-0.228	0.068	4.74
skmeans.5	0.276	26.748	4.37	0.094	-0.363	0.301	0.841	8.873	-2.331	7.976		-0.046	-0.241	0.072	4.681
skmeans.6	0.292	22.651	3.465	0.113	-0.318	0.274	0.857	6.482	-2.013	7.463		-0.051	-0.205	0.071	4.663
skmeans.7	0.268	23.271	4.652	0.11	-0.34	0.243	0.842	7.925	-2.074	11.631		-0.068	-0.205	0.056	3.364
skmeans.8	0.279	19.977	5.702	0.113	-0.316	0.223	0.854	6.881	-1.843	11.565		-0.062	-0.184	0.059	4.59
skmeans.9	0.274	20.092	4.018	0.09	-0.311	0.212	0.851	5.233	-1.831	13.901		-0.069	-0.177	0.055	5.003
skmeans.10	0.242	18.475	3.36	0.074	-0.393	0.221	0.822	4.186	-2.176	11.576		-0.068	-0.222	0.055	7.352
skmeans.11	0.29	19.055	6.271	0.13	-0.253	0.165	0.874	10.214	-1.382	18.638		-0.035	-0.13	0.056	3.368
skmeans.12	0.272	19.535	4.546	0.12	-0.294	0.181	0.855	8.676	-1.654	11.88		-0.027	-0.156	0.062	4.393
skmeans.13	0.295	15.298	4.306	0.088	-0.235	0.135	0.89	2.203	-1.095	24.671		-0.032	-0.11	0.044	7.277
skmeans.14	0.2	17.157	4.768	0.105	-0.478	0.173	0.786	3.95	-2.233	12.245		-0.059	-0.231	0.06	4.825
skmeans.15	0.273	14.471	5.817	0.093	-0.265	0.11	0.877	2.125	-1.09	28.185		-0.106	-0.111	0.043	7.45
skmeans.16	0.19	16.217	4.718	0.11	-0.501	0.155	0.778	4.937	-2.174	11.046		-0.073	-0.228	0.061	5.123
skmeans.17	0.313	15.554	4.224	0.118	-0.131	0.11	0.917	4.429	-0.776	36.195		-0.052	-0.058	0.053	4.308
skmeans.18	0.245	15.061	3.373	0.115	-0.307	0.105	0.848	3.974	-1.293	32.823		-0.077	-0.123	0.05	4.735
skmeans.19	0.188	15.356	4.981	0.095	-0.509	0.157	0.776	4.627	-2.198	11.399		-0.02	-0.232	0.071	6.465
skmeans.20	0.213	14.259	4.801	0.088	-0.412	0.122	0.807	3.295	-1.706	12.238		-0.047	-0.171	0.068	8.009
skmeans.21	0.218	13.454	5.01	0.087	-0.411	0.11	0.817	2.17	-1.541	30.255		-0.054	-0.163	0.053	8.177
skmeans.22	0.19	13.724	2.872	0.111	-0.471	0.12	0.782	2.768	-1.885	13.809		-0.063	-0.19	0.065	4.994
skmeans.23	0.184	13.969	5.12	0.087	-0.485	0.119	0.776	3.467	-1.92	13.978		-0.052	-0.194	0.067	8.006
skmeans.24	0.237	13.282	2.355	0.115	-0.314	0.089	0.847	2.663	-1.193	30.734		-0.063	-0.116	0.058	4.605
skmeans.25	0.23	12.931	3.546	0.082	-0.341	0.087	0.842	2.515	-1.206	24.628		-0.038	-0.123	0.057	9.08
skmeans.26	0.187	12.769	2.924	0.09	-0.486	0.119	0.779	2.337	-1.892	10.411		-0.057	-0.194	0.073	7.363
skmeans.27	0.191	12.937	3.856	0.095	-0.458	0.103	0.788	2.953	-1.705	13.921		-0.051	-0.173	0.07	6.676
skmeans.28	0.169	12.911	3.252	0.087	-0.534	0.127	0.757	2.317	-2.121	8.781		-0.045	-0.217	0.08	7.826
skmeans.29	0.19	11.587	4.909	0.095	-0.437	0.08	0.795	1.598	-1.456	38.46		-0.105	-0.146	0.06	6.717
skmeans.30	0.205	10.771	2.733	0.109	-0.408	0.084	0.813	0.964	-1.37	34.08		-0.085	-0.141	0.059	4.925
skmeans.31	0.194	11.954	4.025	0.09	-0.443	0.088	0.797	1.802	-1.508	29.827		-0.027	-0.155	0.065	7.205
skmeans.32	0.196	12.62	3.08	0.111	-0.389	0.071	0.807	5.821	-1.304	31.275		-0.042	-0.125	0.069	4.721
skmeans.33	0.18	11.55	1.98	0.112	-0.469	0.079	0.783	1.605	-1.51	33.609		-0.057	-0.154	0.068	4.774
skmeans.34	0.193	11.251	3.409	0.108	-0.44	0.082	0.798	1.999	-1.443	32.2		-0.068	-0.148	0.066	5.017
skmeans.35	0.19	11.58	3.259	0.126	-0.447	0.091	0.791	4.857	-1.579	24.132		-0.018	-0.159	0.074	3.636
skmeans.36	0.209	9.958	3.137	0.112	-0.361	0.064	0.831	1.09	-1.096	32.037		-0.08	-0.111	0.066	4.826
skmeans.37	0.172	11.356	3.787	0.115	-0.494	0.078	0.773	4.375	-1.556	30.164		-0.039	-0.16	0.073	4.353
skmeans.38	0.197	10.324	3.341	0.09	-0.412	0.068	0.811	1.867	-1.246	33.362		-0.061	-0.128	0.069	7.11
skmeans.39	0.169	10	3.829	0.081	-0.516	0.095	0.763	3.82	-1.784	9.741		-0.03	-0.182	0.083	8.937
skmeans.40	0.153	9.825	3.264	0.127	-0.558	0.089	0.744	1.1	-1.84	11.825		-0.068	-0.189	0.083	3.653

Tabella A.74: re1 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.489	56.771	4.213	0.029	-0.03	0.256	0.971	10.176	-0.301	6.16	0.037	0.04	-0.021	0.106	48.748
skmeans.3	0.43	51.733	3.919	0.033	-0.053	0.235	0.958	6.315	-0.403	5.517	0.051	0.008	-0.035	0.098	47.328
skmeans.4	0.398	47.512	3.898	0.021	-0.07	0.201	0.943	4.868	-0.51	4.722	0.05	0.025	-0.043	0.101	115.501
skmeans.5	0.394	43.863	2.475	0.021	-0.065	0.183	0.949	4.084	-0.431	4.027	0.051	-0.024	-0.038	0.122	113.223
skmeans.6	0.374	40.253	3.018	0.021	-0.084	0.157	0.934	3.157	-0.514	5.417	0.051	-0.038	-0.045	0.118	111.496
skmeans.7	0.377	37.64	3.87	0.021	-0.063	0.14	0.947	2.397	-0.393	5.216	0.049	-0.061	-0.032	0.07	109.896
skmeans.8	0.379	33.159	3.002	0.021	-0.053	0.132	0.954	1.939	-0.331	5.242	0.047	-0.091	-0.026	0.075	109.488
skmeans.9	0.37	30.659	3.385	0.03	-0.066	0.121	0.945	1.583	-0.376	6.502	0.047	-0.111	-0.031	0.068	54.325
skmeans.10	0.342	30.052	4.314	0.021	-0.12	0.105	0.907	1.917	-0.584	5.777	0.048	-0.079	-0.052	0.062	107.089
skmeans.11	0.34	27.198	4.816	0.033	-0.126	0.103	0.904	1.318	-0.591	6.239	0.048	-0.125	-0.054	0.063	42.794
skmeans.12	0.336	26.25	2.334	0.015	-0.128	0.092	0.901	1.397	-0.575	6.555	0.048	-0.128	-0.052	0.06	211.936
skmeans.13	0.349	23.807	3.711	0.033	-0.095	0.099	0.923	1.085	-0.472	5.929	0.045	-0.115	-0.041	0.072	42.455
skmeans.14	0.33	23.563	3.91	0.015	-0.14	0.085	0.893	1.025	-0.593	6.547	0.047	-0.134	-0.054	0.058	209.816
skmeans.15	0.328	22.506	4.034	0.016	-0.142	0.08	0.891	0.896	-0.588	7.095	0.046	-0.108	-0.053	0.065	208.773
skmeans.16	0.334	21.382	2.109	0.021	-0.117	0.074	0.906	0.849	-0.494	6.998	0.044	-0.127	-0.043	0.067	104.048
skmeans.17	0.314	20.993	3.927	0.032	-0.182	0.08	0.865	0.736	-0.714	7.752	0.046	-0.141	-0.067	0.067	51.585
skmeans.18	0.32	20.317	2.874	0.021	-0.151	0.067	0.883	0.67	-0.574	7.487	0.045	-0.14	-0.051	0.067	102.623
skmeans.19	0.312	18.69	3.85	0.031	-0.174	0.069	0.866	0.738	-0.664	6.061	0.045	-0.144	-0.06	0.062	51.557
skmeans.20	0.317	18.832	3.056	0.023	-0.158	0.063	0.879	0.621	-0.572	6.106	0.044	-0.104	-0.052	0.074	101.905
skmeans.21	0.318	18.167	3.733	0.023	-0.147	0.058	0.883	0.513	-0.534	8.621	0.044	-0.106	-0.047	0.073	101.586
skmeans.22	0.317	17.033	3.908	0.015	-0.162	0.061	0.879	0.457	-0.567	6.553	0.043	-0.133	-0.053	0.058	203.771
skmeans.23	0.31	17.42	4.018	0.023	-0.162	0.054	0.868	0.552	-0.579	9.048	0.043	-0.101	-0.05	0.068	100.476
skmeans.24	0.299	16.505	3.334	0.023	-0.206	0.061	0.843	0.397	-0.719	9.09	0.043	-0.182	-0.065	0.047	100.657
skmeans.25	0.311	16.12	2.203	0.021	-0.167	0.052	0.871	0.425	-0.553	8.305	0.043	-0.126	-0.05	0.063	100.252
skmeans.26	0.306	16.242	3.531	0.021	-0.189	0.055	0.859	0.49	-0.618	13.35	0.045	-0.199	-0.058	0.048	99.218
skmeans.27	0.398	47.512	3.898	0.021	-0.07	0.201	0.943	4.868	-0.51	4.722	0.05	0.025	-0.043	0.101	115.501
skmeans.28	0.394	43.863	2.475	0.021	-0.065	0.183	0.949	4.084	-0.431	4.027	0.051	-0.024	-0.038	0.122	113.223
skmeans.29	0.374	40.253	3.018	0.021	-0.084	0.157	0.934	3.157	-0.514	5.417	0.051	-0.038	-0.045	0.118	111.496
skmeans.30	0.377	37.64	3.87	0.021	-0.063	0.14	0.947	2.397	-0.393	5.216	0.049	-0.061	-0.032	0.07	109.896
skmeans.31	0.379	33.159	3.002	0.021	-0.053	0.132	0.954	1.939	-0.331	5.242	0.047	-0.091	-0.026	0.075	109.488
skmeans.32	0.37	30.659	3.385	0.03	-0.066	0.121	0.945	1.583	-0.376	6.502	0.047	-0.111	-0.031	0.068	54.325
skmeans.33	0.342	30.052	4.314	0.021	-0.12	0.105	0.907	1.917	-0.584	5.777	0.048	-0.079	-0.052	0.062	107.089
skmeans.34	0.34	27.198	4.816	0.033	-0.126	0.103	0.904	1.318	-0.591	6.239	0.048	-0.125	-0.054	0.063	42.794
skmeans.35	0.336	26.25	2.334	0.015	-0.128	0.092	0.901	1.397	-0.575	6.555	0.048	-0.128	-0.052	0.06	211.936
skmeans.36	0.349	23.807	3.711	0.033	-0.095	0.099	0.923	1.085	-0.472	5.929	0.045	-0.115	-0.041	0.072	42.455
skmeans.37	0.33	23.563	3.91	0.015	-0.14	0.085	0.893	1.025	-0.593	6.547	0.047	-0.134	-0.054	0.058	209.816
skmeans.38	0.328	22.506	4.034	0.016	-0.142	0.08	0.891	0.896	-0.588	7.095	0.046	-0.108	-0.053	0.065	208.773
skmeans.39	0.334	21.382	2.109	0.021	-0.117	0.074	0.906	0.849	-0.494	6.998	0.044	-0.127	-0.043	0.067	104.048
skmeans.40	0.314	20.993	3.927	0.032	-0.182	0.08	0.865	0.736	-0.714	7.752	0.046	-0.141	-0.067	0.067	51.585
skmeans.41	0.32	20.317	2.874	0.021	-0.151	0.067	0.883	0.67	-0.574	7.487	0.045	-0.14	-0.051	0.067	102.623
skmeans.42	0.312	18.69	3.85	0.031	-0.174	0.069	0.866	0.738	-0.664	6.061	0.045	-0.144	-0.06	0.062	51.557
skmeans.43	0.317	18.832	3.056	0.023	-0.158	0.063	0.879	0.621	-0.572	6.106	0.044	-0.104	-0.052	0.074	101.905
skmeans.44	0.318	18.167	3.733	0.023	-0.147	0.058	0.883	0.513	-0.534	8.621	0.044	-0.106	-0.047	0.073	101.586
skmeans.45	0.317	17.033	3.908	0.015	-0.162	0.061	0.879	0.457	-0.567	6.553	0.043	-0.133	-0.053	0.058	203.771
skmeans.46	0.31	17.42	4.018	0.023	-0.162	0.054	0.868	0.552	-0.579	9.048	0.043	-0.101	-0.05	0.068	100.476
skmeans.47	0.299	16.505	3.334	0.023	-0.206	0.061	0.843	0.397	-0.719	9.09	0.043	-0.182	-0.065	0.047	100.657
skmeans.48	0.311	16.12	2.203	0.021	-0.167	0.052	0.871	0.425	-0.553	8.305	0.043	-0.126	-0.05	0.063	100.252
skmeans.49	0.306	16.242	3.531	0.021	-0.189	0.055	0.859	0.49	-0.618	13.35	0.045	-0.199	-0.058	0.048	99.218
skmeans.50	0.311	16.12	2.203	0.021	-0.167	0.052	0.871	0.425	-0.553	8.305	0.043	-0.126	-0.05	0.063	100.252

Tabella A.75: tr11 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.532	1.508	5.026	0.003	-0.099	0.273	1.001	301.242	0.08	63.135		0.013	-0.07	0.034	367.738
skmeans.3	0.394	2.239	3.56	0.003	-0.051	0.249	0.984	361.59	-1.05	53.768		-0.145	-0.035	0.062	365.106
skmeans.4	0.254	3.242	3.492	0.003	-0.29	0.254	0.844	637.623	-9.662	61.902		-0.166	-0.182	0.051	360.532
skmeans.5	0.23	2.465	2.942	0.003	-0.274	0.226	0.877	366.277	-7.169	57.733		-0.2	-0.163	0.055	360.394
skmeans.6	0.184	2.847	3.149	0.003	-0.227	0.177	0.874	477.386	-6.573	109.641		-0.289	-0.122	0.058	387.415
skmeans.7	0.145	3.353	2.611	0.003	-0.193	0.152	0.788	516.307	-10.462	511.065		-0.332	-0.098	0.042	382.048
skmeans.8	0.164	2.001	2.942	0.003	-0.248	0.153	0.909	257.249	-4.344	352.792		-0.34	-0.123	0.039	387.562
skmeans.9	0.133	2.365	2.879	0.003	-0.173	0.126	0.838	295.391	-7.257	492.524		-0.356	-0.08	0.037	383.043
skmeans.10	0.144	2.236	2.069	0.003	-0.213	0.129	0.914	332.711	-3.789	233.741		-0.257	-0.098	0.05	381.914
skmeans.11	0.135	2.243	3.408	0.003	-0.153	0.108	0.948	294.538	-2.125	420.223		-0.296	-0.066	0.045	379.796
skmeans.12	0.113	2.441	2.829	0.003	-0.118	0.089	0.908	291.785	-3.498	384.438		-0.288	-0.047	0.047	474.606
skmeans.13	0.129	1.779	2.918	0.003	-0.185	0.094	1.023	170.827	0.882	204.135		-0.219	-0.074	0.062	480.71
skmeans.14	0.102	2.205	2.593	0.003	-0.177	0.089	0.858	241.461	-5.307	626.384		-0.292	-0.069	0.05	374.126
skmeans.15	0.109	1.949	3.274	0.003	-0.17	0.083	0.946	164.905	-1.934	483.805		-0.203	-0.064	0.053	473.902
skmeans.16	0.107	1.899	2.457	0.003	-0.213	0.087	0.93	197.002	-2.538	365.321		-0.282	-0.081	0.058	472.491
skmeans.17	0.107	1.689	2.834	0.003	-0.245	0.085	0.95	129.903	-1.755	588.904		-0.3	-0.09	0.046	375.388
skmeans.18	0.097	1.972	3.688	0.003	-0.141	0.068	0.944	156.482	-1.844	565.376		-0.249	-0.049	0.051	391.499

Tabella A.76: tr12 SKmeans

Partizioni	C	CH	DB	D	G	G_p	MCR	PBM	PB	RT	RL	S	T	WG	XB
skmeans.2	0.439	2.448	3.515	0.004	-0.315	0.328	0.784	756.621	-16.668	30.301	0.084	0.077	-0.223	0.269	367.323
skmeans.3	0.344	2.134	3.065	0.003	-0.361	0.33	0.774	685.325	-16.827	98.911	0.093	-0.155	-0.251	0.066	413.237
skmeans.4	0.322	1.574	2.75	0.003	-0.24	0.27	0.928	451.915	-4.726	243.597	0.085	-0.209	-0.159	0.058	482.45
skmeans.5	0.258	1.435	2.897	0.003	-0.185	0.207	1.006	310.932	0.321	214.699	0.081	-0.227	-0.109	0.058	440.224
skmeans.6	0.23	1.39	3.142	0.003	-0.134	0.174	1.04	274.309	2.133	230.69	0.079	-0.233	-0.074	0.076	478.982
skmeans.7	0.191	1.533	2.921	0.003	-0.203	0.161	0.987	363.847	-0.676	222.744	0.083	-0.255	-0.105	0.064	475.527
skmeans.8	0.168	1.814	3.235	0.003	-0.149	0.144	0.93	872.921	-3.426	215.1	0.094	-0.292	-0.074	0.057	470.244
skmeans.9	0.142	1.748	2.62	0.003	-0.255	0.14	0.882	612.035	-5.46	235.588	0.091	-0.248	-0.12	0.075	468.28
skmeans.10	0.152	1.144	2.75	0.001	-0.234	0.133	0.965	203.021	-1.601	336.563	0.078	-0.326	-0.109	0.051	3421.312
skmeans.11	0.134	1.143	2.818	0.003	-0.241	0.118	0.948	192.718	-2.187	225.195	0.076	-0.292	-0.105	0.07	471.953
skmeans.12	0.117	1.295	2.654	0.003	-0.315	0.11	0.919	255.063	-3.235	352.808	0.082	-0.287	-0.129	0.068	467.681
skmeans.13	0.126	1.119	3.556	0.003	-0.275	0.111	0.96	185.774	-1.616	277.943	0.076	-0.225	-0.115	0.082	468.843
skmeans.14	0.094	1.509	2.961	0.003	-0.297	0.1	0.803	260.17	-7.601	428.25	0.088	-0.244	-0.117	0.06	459.665
skmeans.15	0.1	1.179	3.117	0.003	-0.218	0.085	0.911	151.039	-3.241	499.125	0.075	-0.251	-0.082	0.064	464.108
skmeans.16	0.108	1.225	2.581	0.003	-0.271	0.092	0.957	258.187	-1.589	718.643	0.08	-0.222	-0.103	0.069	461.289

Appendice B

Routine in linguaggio R

```
####pacchetti da caricare####
#########################
library(cluster)
library (skmeans)
library(clusterCrit)
library(tm)
library(koRpus)

####Procedura di esempio####
##########################
source('partizionikmeans.r')
dataset_partizioni<-partizionikmeans(dataset,kmin,kmax)
source('partizioniupgma.r')
dataset_partizioni<-partizioniupgma(dataset,kmin,kmax)
source('partizioniward.r')
dataset_partizioni<-partizioniward(dataset,kmin,kmax)
```

```
save(dataset_partizioni, file="dataset_partizioni.RData")
#Calcolo degli indici e dei ranghi
source('indici.r')
source('ranghi.r')
val_dataset<-t(indices(dataset,dataset_partizioni))
#salvare le matrici con i valori degli indici
save(val_dataset, file="val_dataset.Rdata")
rank_dataset<-ranghi(val_dataset)
save(rank_dataset, file="rank_dataset.Rdata")
risultati_dataset<-list(val_dataset,rank_dataset)
write.table(ris_dataset, file="ris_dataset.csv", dec=",")

####Partizioni Kmeans####
######################
partizionikmeans<-function(dati,kmin,kmax){
dati<-data.matrix(dati)
partizioni<-matrix(0,nrow(dati),(kmax-kmin+1))
k<-kmin
i=1
while (kmin<=kmax)
{kmeans <- kmeans(dati,kmin)
partizioni[,i]<-(as.integer(kmeans$cluster))
i=i+1
kmin=kmin+1
}
nomi1<-paste("kmeans",k:kmax, sep=".")
colnames(partizioni)<-nomi1
```

```
partizioni}

###Funzione per calcolare gli indici di validazione
interna, per ciascuna partizione#######
################################
indici<-function(data,partition){
library(clusterCrit)
data<-data.matrix(data)
#funzioni Pacchetto ClusterCrit
#getCriteriaNames returns the available clustering criteria
names.
#intCriteria calculates various internal clustering
validation or quality criteria.
#intCriteria(matrix of observations, partition vector,
vector containing the names of the indices to compute)
index<-c("C_index","Calinski_Harabasz","Davies_Bouldin",
"Dunn","Gamma","G_plus","McClain_Rao","PBM",
"Point_Biserial","Ray_Turi","Ratkowsky_Lance","Silhouette",
"Tau","Wemmert_Gancarski","Xie_Beni")
n.index<-length(index)
nomi<-list(index,colnames(partition))
matrice<-matrix(0,n.index,ncol(partition),dimnames=nomi)
for (j in 1:(ncol(partition)))
{prova<-intCriteria(data, as.integer(partition[,j]), index)
for (i in 1:n.index)
{matrice[i,j]<-prova[[i]]}
}
```

```
matrice<-round(matrice,3)
}

####Calcolo dei ranghi ####
#######################
ranghi<-function(dati)
{
matrice_ranghi<-matrix(0,nrow(dati),ncol(dati))
colnames(matrice_ranghi)<-colnames(dati)
rownames(matrice_ranghi)<-rownames(dati)
criterio<-c("min","max","min","max","min","min","min",
"max","min","min","max","max","min","max","min")
names(criterio)<-c("C_index", "Calinski_Harabasz",
"Davies_Bouldin","Dunn","Gamma","G_plus","McClain_Rao",
"PBM","Point_Biserial","Ray_Turi","Ratkowsky_Lance",
"Silhouette","Tau","Wemmert_Gancarski", "Xie_Beni")
for (j in 1:ncol(dati)){
if (criterio[j]=="min"){
matrice_ranghi[,j]<-rank(dati[,j],ties.method="min")}
else
{matrice_ranghi[,j]<-rank(-(dati[,j]),ties.method="min")}
}
matrice_ranghi
}

####Partizioni Average-Linkage####
#############################
```

```
partizioniupgma<-function(dati,kmin,kmax){
dati<-data.matrix(dati)
dissimilarity<-dist(dati)
partizioni<-matrix(0,nrow(dati),(kmax-kmin+1))
k<-kmin
i=1
upgma <- agnes(dissimilarity)
while (kmin<=kmax)
{partizioni[,i]<-cutree(upgma,kmin)
i=i+1
kmin=kmin+1
}
nomi1<-paste("upgma",k:kmax, sep=".")
colnames(partizioni)<-nomi1
partizioni}

#####Partizioni Ward####
######################
partizioniward<-function(dati,kmin,kmax){
dati<-data.matrix(dati)
dissimilarity<-dist(dati)
partizioni<-matrix(0,nrow(dati),(kmax-kmin+1))
k<-kmin
i=1
ward<- hclust(dissimilarity, method="ward")
while (kmin<=kmax)
{partizioni[,i]<-cutree(ward,kmin)
```

```
i=i+1
kmin=kmin+1
}
nomi1<-paste("ward",k:kmax, sep=".")
colnames(partizioni)<-nomi1
partizioni}

#####Importare i corpora######
#########################
#cambiare la directory di lavoro
setwd("~/Documents/DATASET/.....")
#getwd() recupera il percorso della directory di lavoro
e carica tutti i file di testo presenti
corpus<-Corpus(DirSource(getwd()))
corpus
#vedere il contenuto dell'oggetto corpus
inspect(corpus)
#vedere le funzioni che consentono di trasformare il corpus
getTransformations()
#"as.PlainTextDocument" "removeNumbers"  "removePunctuation"
#"removeWords"          "stemDocument"    "stripWhitespace"
#stopwords("english")

####Pretrattamento####
###################
corpus<-tm_map(corpus, removeNumbers)
corpus<-tm_map(corpus, removePunctuation)
```

```
corpus<-tm_map(corpus, tolower)
corpus<-tm_map(corpus, stripWhitespace)
corpus<-tm_map(corpus, removeWords, stopwords("english"))
lemma<-list()
for (i in 1:n.documents)
{
tagged.text <- treetag(corpus[[i]], treetagger="manual",
lang="en", TT.options=list(path="/Users/
Documents/tree-tagger",preset="en"), format = "obj")
lemma[[i]]<-tagged.text@TT.res$lemma
}
corpus_l<-Corpus(VectorSource(lemma))
dtm <- DocumentTermMatrix(corpus_l)

####Partizioni Skmeans####
######################
partizioniskmeans<-function(dati,kmin,kmax){
library(skmeans)
partizioni<-matrix(0,nrow(dati),(kmax-kmin+1))
k<-kmin
i=1
while (kmin<=kmax)
{skmeans <- skmeans(dati,kmin)
partizioni[,i]<-(as.integer(skmeans$cluster))
i=i+1
kmin=kmin+1
}
```

```
nomi1<-paste("skmeans",k:kmax, sep=".")
colnames(partizioni)<-nomi1
partizioni}
```

Bibliografia

[1] Ackerman M., Ben-David S. (2009) Clusterability: A theoretical study, In *Proceedings of 12th International Conference on Artificial Intelligence and Statistics* 1-8.

[2] Balbi S., Misuraca M. (2005) Pesi e Metriche nell'Analisi dei Dati Testuali, *Quaderni di Statistica*, Liguori, 7, 55–68.

[3] Baker F.B., Hubert L.J. (1975) Measuring the power of hierarchical cluster analysis, *Journal of the American Statistical Association*, 70, 31-38.

[4] Bandyopadhyay S., Pakhira M.K., Maulik U. (2004) Validity index for crisp and fuzzy clusters, *Pattern Recognition*, 37, 487-501.

[5] Bellman R. (1961). Adaptive Control Processes. A Guided Tour. Princeton University Press

[6] Benzécri J. P. (1981) Pratique de l'Analyse des Données, Vol. 3, *Linguistique et Lexicologie*, Dunod, Paris.

[7] Benzécri J.P. (1982) *Histoire et Préhistoire de l'Analyse des Données*, Bordas, Dunod, Paris.

[8] Beyer K., Goldstein J., Ramakrishnan R., Shaft U. (1999) When is "nearest neighbor" meaningful?, In *Proceedings of the 7th International Conference on Database Theory*, Springer Berlin Heidelberg, 217-235.

[9] Bezdek J.C., Pal N.R. (1998) Some new indexes of cluster validity, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 28(3), 301-315.

[10] Bishop C. (1995) *Neural Networks for Pattern Recognition*, New York: Oxford Univ. Press.

[11] Bolasco S. (1990) Sur différentes stratégies dans une analyse des formes textuelles: une expérimentation à partir de données d'enquête, In *Bécue M., Lebart L., Rajadell N. editors, Actes des Premières journées JADT*, UPC, Barcelone, (1992), 69-88.

[12] Bolasco S. (1999) *Analisi multidimensionale dei dati*, Carocci ed., Roma.

[13] Calinski R.B., Harabasz J. (1974) A dendrite method for cluster analysis, *Communications in Statistics*, 3, 1-27.

[14] Cohen W.W. (1996) Learning rules that classify e-mail, In *AAAI spring symposium on machine learning in information access*, 18, 25.

[15] Cook T.D., Campbell D.T., Day A. (1979) *Quasi-experimentation: Design analysis issues for field settings*, 351, Boston: Houghton Mifflin.

[16] Croft W.B. (1977) Clustering large files of documents using the single-link method, *Journal of the American Society of Information Science*, 28, 341–344.

[17] Davies D.L., Bouldin D.W. (1979) A cluster separation measure, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2), 224-227.

[18] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshmanet R. (1990) Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41(6), 39-407.

[19] De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993) *Lessico di frequenza dell'italiano parlato*, EtasLibri, Milano.

[20] Dhillon I.S., Modha D.S. (2001) Concept decompositions for large sparse text data using clustering, *Machine Learning*, 42(1-2), 143-175.

[21] Dixon, L.C.W., Szegö, G.P. (1978) The global optimization problem: an introduction, *Towards global optimization*, 2, 1-15.

[22] Dubes R., Jain A.K. (1979) Validity studies in clustering methodologies, *Pattern Recognition*, 11(4), 235-254.

[23] Dunn† J.C. (1974) Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics*, 4(1), 95-104.

[24] Eckart C., Young G. (1936) The approximation of one matrix by another of lower rank, *Psychometrika*, 1, 211-218.

[25] Efron B., Efron B. (1982) *The jackknife, the bootstrap and other resampling plans*, Philadelphia: Society for industrial and applied mathematics.

[26] Everitt B.S. (1979) Unresolved problems in cluster analysis *Biometrics*, 169-181.

[27] Everitt B.S., Dunn G. (1991) *Applied Multivariate Data Analysis*, Edward Arnold, London.

[28] Everitt B.S., Landau S., Leese M. (2001) *Cluster Analysis*, Fourth edition, Arnold.

[29] Fowlkes E.B., Mallows C.L. (1983) A method for comparing two hierarchical clusterings, *Journal of the American statistical association*, 78(383), 553-569.

[30] Gordon, A.D. (1999) *Classification*, Chapman and Hall/CRC, London.

[31] Gower, J.C. (1971) A general coefficient of similarity and some of its properties, *Biometrics*, 857-871.

[32] Greenacre M.J. (1984) *Theory and Application of Correspondence Analysis*, Academic Press, London.

[33] Gross M. (1968) *Grammaire transformationnelle du français: Syntaxe du verbe*, Larousse, Paris.

[34] Halkidi M., Vazirgiannis M., Batistakis Y. (2000) Quality scheme assessment in the clustering process, In *Principles of Data Mining and Knowledge Discovery*, Springer Berlin Heidelberg, 265-276

[35] Halkidi M., Batistakis Y., Vazirgiannis M. (2001) On clustering validation techniques, *Journal of Intelligent Information Systems*, 17(2-3), 107-145.

[36] Hinneburg, A., Aggarwal, C. C., Keim, D. A. 2000. What is the nearest neighbor in high dimensional spaces?, In *Proceedings of the 26th International Conference on Very Large Data Bases.*

[37] Hopkins B. (1954) A new method of determining the type of distribution of plant individuals, *Annals of Botany*, 18, 213-226.

[38] Hubert L., Schultz J. (1976) Quadratic assignment as a general data-analysis strategy, *British Journal of Mathematical and Statistical Psychologie*, 29, 190-241.

[39] Hubert L., Arabie P. (1985) Comparing partitions, *Journal of classification*, 2(1), 193-218.

[40] Iezzi D.F. (2012) A new method for adapting the k-means algorithm to text mining, *Statistica Applicata*, 22, 69-80.

[41] Jaccard P. (1912) The distribution of the flora in the alpine zone, 1, *New phytologist*, 11(2), 37-50.

[42] Jain A., Dubes R., *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[43] Jain A., Murty M., Flynn P. (1999) Data clustering: A review, *ACM computing surveys (CSUR)*, 31(3),264–323.

[44] Jain A. K. (2010) Data clustering: 50 years beyond K-means, *Pattern recognition letters*, 31(8), 651-666.

[45] James. M. (1985) *Classification Algorithms*, Wiley Interscience.

[46] Kaufman L., Rousseeuw P.J. (2009) *Finding groups in data: an introduction to cluster analysis*, 344, John Wiley & Sons.

[47] Kendall, M. (1938) A New Measure of Rank Correlation, *Biometrika*, 30, (1–2) 81–89.

[48] Kleinberg J. (2002) An impossibility theorem for clustering, in *Proc. 2002 Conf. Advances in Neural Information Processing Systems*, 15, 463–470.

[49] Krippendorff K. (1980) *Content Analysis: An Introduction to Its Methodology*, Sage Publications, Beverly Hills, CA.

[50] Kusiak A. (2001) Feature transformation methods in data mining, *Electronics Packaging Manufacturing, IEEE Transactions on*, 24(3), 214-221.

[51] Lam B.S., Yan H. (2005) A new cluster validity index for data with merged clusters and different densities, In *Systems, Man and Cybernetics, 2005 IEEE International Conference*, 1, 798-803

[52] Lam W., Ho. C.Y. (1998) Using a generalized instance set for automatic text categorization, *ACM SIGIR Conference*.

[53] Larsen B., Aone C. (1999) Fast and effective text mining using linear-time document clustering, In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 16-22.

[54] Lebart L., Salem A. (1988) *Analyse statistique des données textuelles*, Dunod, Paris.

[55] Lebart L., Salem A. (1994) *Statistique textuelle*, Dunod, Paris.

[56] Liu Y., Li Z., Xiong H., Gao X., Wu J. (2010) Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, IEEE, 911-916.

[57] von Luxburg U. (2009) Clustering stability: An overview, *Foundations and Trends in Machine Learning*, 2 (3), 235–274.

[58] MacQueen J. (1967) Some methods for classification and analysis of multivariate observations, In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*,1(14), 281-297.

[59] McCallum A., Nigam K. (1998) A comparison of event models for naive bayes text classification, In *AAAI-98 workshop on learning for text categorization*, 752, 4-48.

[60] McClain J.O., Rao V.R. (1975) Clustisz: A program to test for the quality of clustering of a set of objects, *Journal of Marketing Research*, 12, 456-460.

[61] Milligan G.W. (1981) A monte carlo study of thirty internal criterion measures for cluster analysis, *Psychometrika*, 46, (2), 187-199.

[62] Milligan G.W., Cooper M C. (1985) An examination of procedures for determining the number of clusters in a data set, *Psychometrica*, 50,159-179.

[63] Oren N. (2002) Reexamining tf.idf based information retrieval with genetic programming, In *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, South African Institute for Computer Scientists and Information Technologists, 224-234.

[64] Rand W.M.(1971) Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical association*, 846-850.

[65] Ratkowsky D.A., Lance G.N. (1978) A criterion for determining the number of groups in a classification, *Australian Computer Journal*, 10, 115-117.

[66] Ray S., Turi R.H. (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation, In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, 137-143.

[67] Reinert M. (1986) Un logiciel d'analyse lexicale: ALCESTE. *Les Cahiers de l'analyse des données*, XI(4), 471-484.

[68] Rohlf F.J. (1974) Methods of comparing classifications, *Annual Review of Ecology and Systematics*, 5, 101-113.

[69] Rousseeuw P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, 20, 53-65.

[70] Salem A. (1987) Pratique des segments répétés, *Essai de statistique textuelle*, Klincksieck, Paris.

[71] Salton G., Buckley C. (1988) Term-weighting approaches in automatic text retrieval, *Information Processing e Management*, 24(5), 513-523.

[72] Shannon C.E. (2001) A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), 3-55.

[73] Sharma S. (1996) *Applied multivariate techniques*, John Wiley & Sons, Inc.

[74] Silvestri L.G., Hill L.R. (1964) Some problems of the taxometric approach, *Phenetic and Phylogenetic Classification, Syst. Ass. Pub*, 6, 87-103.

[75] Sneath P.H., Sokal R.R. (1973) *Numerical taxonomy. The principles and practice of numerical classification.*

[76] Sokal R.R., Sneath P.H. (1963) *Principles of numerical taxonomy*, Principles of numerical taxonomy.

[77] Spärck Jones K. (1972) A statistical interpretation of term specificity and its application in retrieval, In *Journal of Documentation*, 28(1), 11-21.

[78] Steinbach M., Karypis G., Kumar. V. (2000) A comparison of document clustering techniques, *Technical report, Department of Computer Science and Engineering*, University of Minnesota.

[79] Tan P.N., Steinbach M., Kumar V. (2006) *Introduction to data mining*, Vol.1, Boston: Pearson Addison Wesley.

[80] Tibshirani R., Walther G., Hastie T. (2001), Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.

[81] Tryon R.C. (1939) *Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*, Edwards brother, Incorporated, lithoprinters and publishers.

[82] Tyron R.C., Bailey D.E. (1970) *Cluster Analysis*, McGraw-Hill.

[83] Vendramin L., Campello R.J., Hruschka E.R. (2009) On the Comparison of Relative Clustering Validity Criteria. In *SDM*, 733-744.

[84] Voorhees E.M. (1986) Implementing Agglomerative Hierarchical Clustering for use in Information Retrieval, *Technical Report TR86–765*, Cornell University, Ithaca, NY.

[85] Ward Jr J.H. (1963) Hierarchical grouping to optimize an objective function, *Journal of the American statistical association*, 58(301), 236-244.

[86] Wang W., Wang C., Cui X., Wang A. (2008) Fuzzy C-Means Text Clustering with Supervised Feature Selection, In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference*, 1, 57-6.

[87] Wemmert C., Gançarski P., Korczak J. (2000) A collaborative approach to combine multiple learning methods, *International Journal on Artificial Intelligence Tools (World Scientific)*, 9(1), 59–78.

[88] Willett P. (1988) Recent Trends in Hierarchical Document Clustering: A Critical Review, *Information Processing and Management*, 24(5), 577–597.

[89] Xie X.L., Beni G. (1991) A validity measure for fuzzy clustering, *IEEE Transactions on pattern analysis and machine intelligence*, 13(8), 841-847

[90] Yule G.U. (1944) *A statistical study of vocabulary*, Cambridge, Cambridge Univ. Press..

[91] Zhao Y., Karypis G. (2002) Evaluation of hierarchical clustering algorithms for document datasets, In *Proceedings of the eleventh international conference on Information and knowledge management*, ACM, 515-524.

[92] Zipf G.K. (1935) *The psychobiology of language. An introduction to dynamic philology*, Houghton-Mifflin, Boston.