Contents

Elenco delle Tabelle IX

Introduzione 1

1 Knowledge discovery and data mining 3
1.1 Introduction 3
1.2 Data mining 6
1.2.1 Functionalities of dat26(mining)-583(.)-500(.)-500(.)-500(.)-500

14

2.8.1 Partition algorithm 34
2.8.2 Eclat algorithm 35
2.8.3 FP-growth algorithm 37
2.9 Extentions of association rules mining framework . . . 39
2.9.1 Quantiott(w)218(lgo)ndwc(ion)egatd217esng2.9 582500(.)-500(.)-500(.)-50

4.6 Examples of application 70

List of Figures

1.1 Scheme of the knowledge discovery and data mining . 7

2.1 *Lattice representation of the itemset search space* . . . 24
2.2 *Lattice representation of the association rules search* . 28

Pr

Il primo capitolo.....

Nel secondo capitolo....
Nel terzo capitolo.....
Nel quarto capitolo.....
La tesi si conclude...

Chapter 1

Knowledge discovery and data mining

1.1 Introduction

In the present age of the information society each individual, family or

tion. From business transactions and scientific data, to satellite pictures, text reports and military intelligence, information retrieval is not enough anymore for decision-making. Dealing with huge collections of data, there are new needs to help managerial decisions. These

patterns.

1.3.1 Statistical tools for data mining

In any data mining process a core phase is the selection of the more

of one or more variables through the analysis of all the remaining considered variables. The predictive task is achieved exploiting *neural networks* and *decision trees*

Dr. Strangelove also buys *The Great Escape*.

The mining of association rules is not only applied in market basket analysis but has a wider field of application: *text mining*, *image mining*, *multimedia mining* and *remote sensed imagery*. However, the first and best known field of application of association rule mining is the analysis of transactional data bases.

The basic record of a transactional data base is the transaction, and it represents, as a classic example, the collection of the products purchased by a single costumer of a suermarket in a row. The whole data base D can be seen as a binary matrix with transactions on rows and products (items) of the considered set I on columns. Furthermore, each transaction is characterized by a transaction identifier (*tid*).

The data base layout can be *horizontal* and in this case D is considered a transactions collection, rows of the binary matrix; similarly, when using a *horizontal* layout, the data base is a co102-381(a)-395(sk6e[().)]Titems,)-414(t

way. The overview continues in section 2.9 where are references of the contributions aiming to extend the applicability of AR: in particular

$$support(X, D) := |cover(X, D)|. \tag{2.3}$$

The *frequency* of an itemset X in

AR	*sup.*	*conf.*
beer → chips	0.50	1
beer → wine	0.25	0.50
chips → beer	0.50	0.66
pizza → chips	0.25	0.50
pizza → wine	0.25	0.50
wine → beer	0.25	0.50
wine → chips	0.25	0.50
wine → pizza	0.25	0.50
beer, chips → wine	0.25	0.50
beer, wine → chips	0.25	1
chips, wine → beer	0.25	1
beer → wine, chips	0.25	0.50
wine → beer, chips	0.25	0.50

Table 2.2: *Association rules in D*

part of the present chapter the two phases of the AR mining strategy are described in depth.

2.3 Itemset mining

Discovering frequent itemsets in a data base is a problem hard to solve,

sets. The formal definition is:

Definition 3 (Candidate itemset). *Given aansaction dataase D*

$$Bd^{+}(F) := \{X \quad I \mid \quad Y \quad X : Y \quad F \qquad Z \quad X : Z \ / \ F\}.$$

and those itemsets in $Bd(F)$ that are not in F are called *negative* $Bd^{-}(F)$:

$$Bd^{-}(F) := \{X$$

2.4 Assoc

The number of

This means that the confidence of a rule monotone decreases as the

Algorithm 1 Apriori algorithm: frequen itemset mining

Figure 2.2: *Lattice representation of the association rules search*

finally added to F_k, thus are frequent itemsets. A further aspect to no8(t)ice is that the set o(t)f the candidate itemse8(t)s that turn o(t)ut to b infrequent equals to Bd^-

Algorithm 2 Apriori: association rule mining

Input: D, σ, γ;
Output: $R(D, \sigma, \gamma)$;

1: Compute $R(D, \sigma)$;
2: $R := \{\}$
3: **for all** $I \quad F$ **do**
4:

the number of records in Cs_k becomes more and more smaller as the

Proposition 4. *Let $X, Y, Z \quad I$ be itemsets.*

$$support(X \quad Y \quad Z) \geq support(X \quad Y) + support(X \quad Z) - suppot(X)$$

to be read in main memory. In real world applications, the huge size

and ignored in the rest of the procedure.

Algorithm 3 Eclat

Input

ported in algorithm 3 and 4 respectively. The advantage of the FP-growth over Eclat is that each linked list that starts from an item in the header table, representing the cover of that item, is stored in a compressed form. The problem is that, in order to obtain the compressed storage form, the procedure needs to build up and maintain aemcomplexemdataemstructure, performing dereferencing, while Eclat performs just simple interactions. In addiction, the cover of an item in Eclat is implemented using an array of transaction identifiers; in FP-growth the cover of an item is compressed using the linked list starting from its node list in the header table. Every node in the linked list needs the storage (of)6]TJ04i91(the4(an)-3eds)-236(0(ite(tem)ra)292(h)1(e)ouct

confidence; this is not always a desirable objective, since it can often be useful to mine rules with an high confidence and with no necessary high support. High confidence rule mining is useful,for example, to detect the buying behaviors of a little segment of the supermarket customers (a market nich2(a)-). Some of the proposals in this AR mining research issues are reported in the following of this section.

of computer I/O operations, but it is possible to roughly classify them in:

- rule *quantity* problem: too many rules are being generated by the process;
- rule *quality* problem: not all the generated rules are interessing.

It is useful to rl t3

the distribution under a null hypothesis that, when testing for de-

where $Cov(A, B) = E(AB) - E(A)E(B)$. The correlation coe□cient has a range between -1 and $+1$, when $\rho_{A,B} = 0$, then the two variables are independent. For binary variables, that is the classic data□structure in association rule mining, the standard deviation is $\sigma_A =$

On the basis of the above considerations, equation 2.16 can be rewritten as follows:

$$\begin{aligned}\lambda_A &= \frac{\sum_j max_k P(A_k, B_j) - max_k P(A_k)}{1 - max_k P(A_k)} \\ &= \frac{\sum_j max_k f_{jk} - max_k f_{+k}}{N - max_k f_{+k}}.\end{aligned} \tag{2.18}$$

This form of the λ-coe□cient measures the predictive association of the rule $B \rightarrow A$. It is possible to cosider a sytric version of such

association between variables A and B the maximum mutual information ratio can be used:

$$S(A, B) = \frac{H(A) + H(B) - H(A, B)}{}$$

	B

$$\varphi \approx I \sqrt{\frac{f_{1+} f_{+1}}{N^2}}$$

$$= \frac{N f_{11}}{}$$

Chapter 3

Clustering in data mining

3.1 Introduction

The general aim of clustering is to determine groups of objects, with respect to an observed data set, such that the homogeneity within each group is maximized and the homogeneity between di□erent groups is minimized. Cluster analysis is used in a wide number of di□erent

data. Furthermore, in real world the application the attributes to consider are many and of di□erent nature, then procedures able to deal with di□erent kinds of attributes are required. In addiction, in case of large data bases it is quite subjective to assign a priori any parameter to a clustering algorithm: for example, the choice of the number of clusters to generate strongly a□ects quality and e□ectiveness of the results. Clustering results does not have to be influenced neither by the input order of data and the presence of outliers and missing data. A last but not list feature required to a data mining clustering procedure is the interpretability and usability of the results: a user friendly presentation of the results is crucial in extracting useful knowledge from data.

3.2.1 Categories of clustering methods

There are many proposals in the literature aiming at clustering, and the di□erences among the procedures regard the nature of the data they deal with as well as the purpose of the procedure application. In wide meaning, the clustering methods can be in di□erent categories:

In most of the real world applications, a transactional data base contains thousands and thousands of transactions observed on hundreds of di□erent attributes. In order to deal with very large and high dimensional data base, di□erent ad-h33919[tstr(ing)-291prh3396(c(e)-1duerv)27(ed4g)-2h

Chapter 4

Exploratory association analysis

4.1 Introduction

4.2 Steps of the exploratory study of the association

4.3 Data structures

Bibliography