Association Rule Mining: Applications in Various Areas

Akash Rajak and Mahendra Kumar Gupta

Krishna Institute of Engineering & Technology, 13 K.M. Stone, Delhi-Merrut Highway,
Ghaziabad-201206, (U.P.)

ABSTRACT

This paper presents the various areas in which the association rules are applied for effective decision making. Association rule mining seeks to discover associations among transactions encoded in a database. It can be used to improve decision making in a wide variety of applications such as: market basket analysis, medical diagnosis, bio-medical literature, protein sequences, census data, logistic regression, fraud detection in web, CRM of credit card business etc.

Keywords: *Data mining, association rule, market basket analysis, protein sequences, logistic regression.*

1. INTRODUCTION

Association rules have been broadly used in many applications domains for finding pattern in data. The pattern reveals combinations of events that occur at the same time. One of the best domain is business field, where discovering of pattern or association helps in effective decision making and marketing. Other areas where association rule mining can be applied, are finding pattern in biological databases, market basket analysis of library circulation data, to study protein composition, to study population and economic census etc.

Recent studies have shown that there are various algorithms for finding association rule. One of the best known algorithm is apriori algorithm. However the complexity and performance of mining algorithms is subject to research area, as they have to mine a larger set of data items. i.e. most of the study are based on how to simplify association rule and to improve the algorithm performance.

2. BACKGROUND

Let us introduce the foundation of association rule and their significance. Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).

Support: The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. (The support is sometimes expressed as a percentage of the total number of records in the database.)

Confidence: Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

Lift: Lift is nothing but the ratio of confidence to expected confidence. Lift is a value that gives us information about the increase in probability of the "then" (consequent) given the "if" (antecedent) part.

3. APPLICATION AREAS

The various application areas in which association rules can be applied for extracting useful information from the huge dataset are:

3.1 Market basket analysis

A typical and widely-used example of association rule mining is market basket analysis. For example, data are collected using bar-code scanners in supermarkets. Such ‘market basket’ databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns [7].

For example, if a supermarket database has 100,000 point-of-sale transactions, out of which 2,000 include both items A and B and 800 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 800 transactions (alternatively 0.8% = 800/100,000) and a confidence of 40% (=800/2,000).

One way to think of support is that it is the probability that a randomly selected transaction from the database will contain all items in the antecedent and the consequent, whereas the confidence is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.

Now days every product comes with bar code. The software supporting these barcode based purchasing/ordering systems produces vast amounts of sales data, typically captured in “baskets” (records in which the items purchased by a given consumer at a given time are grouped together). This data was quickly recognized by the business world as having immense potential value in marketing. In particular, commercial organizations are interested in discovering “association rules” that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This “market basket analysis” result can then be used to suggest combinations of products for special promotions or sales, devise a more effective store layout, and give insight into brand loyalty and co-branding.

Market basket can be defined as collection of items purchased by a customer in a single transaction (e.g. supermarket, web) Association rules are used for pattern discovery, each rule has form: A->B, or Left -> Right. For example: "70% of customers who purchase 2% milk will also purchase whole wheat bread."

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B, i.e.

Support = Probability (A and B)

Support = (# of transactions involving A and B) / (total number of transactions).

Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A, i.e. Confidence = Probability (B if A) = P (B/A)

Confidence = (# of transactions involving A and B) / (total number of transactions that have A). Example:

Table 1. An example

Customer	**Item purchased**	**Item purchased**
1	pizza	beer
2	salad	soda
3	pizza	soda
4	salad	tea

If A is "purchased pizza" and B is "purchased soda" then

Support = P (A and B) = ¼

Confidence = P (B / A) = ½

Confidence does not measure if the association between A and B is random or not.

3.2 Medical diagnosis

Applying association rules in medical diagnosis can be used for assisting physicians to cure patients. The general problem of the induction of reliable diagnostic rules is hard because theoretically no induction process by itself can guarantee the correctness of induced hypotheses [5].

Practically diagnosis is not an easy process as it involves unreliable diagnosis tests and the presence of noise in training examples. This may result in hypotheses with unsatisfactory prediction accuracy which is too unreliable for critical medical applications [2].

Serban [5] has proposed a technique based on relational association rules and supervised learning methods. It helps to identify the probability of illness in a certain disease. This interface can be simply extended by adding new symptoms types for the given disease, and by defining new relations between these symptoms.

3.3 Protein sequences

Proteins are important constituents of cellular machinery of any organism. Recombinant

DNA technologies have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes [1].

Proteins are sequences made up of 20 types of amino acids. Each protein has a unique 3-dimensional structure, which depends on amino-acid sequence; slight change in sequence may change the functioning of protein. The heavy dependence of protein functioning on its amino acid sequence has been a subject of great anxiety.

Lot of research has gone into understanding the composition and nature of proteins; still many things remain to be understood satisfactorily. It is now generally believed that amino acid sequences of proteins are not random.

Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra [9] have deciphered the nature of associations between different amino acids that are present in a protein. Such association rules are desirable for enhancing our understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occurring in proteins. Knowledge of these association rules or constraints is highly desirable for synthesis of artificial proteins.

3.4 Census data

Censuses make a huge variety of general statistical information on society available to both researchers and the general public [3]. The information related to population and economic census can be forecasted in planning public services(education, health, transport, funds) as well as in public business(for setup new factories, shopping malls or banks and even marketing particular products).

The application of data mining techniques to census data and more generally to official data, has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society [4]. On the other hand, it is not undemanding and requires exigent methodological study, which is still in the preliminary stages.

3.5 CRM of credit card business

Customer Relationship Management (CRM), through which, banks hope to identify the preference of different customer groups, products and services tailored to their liking to enhance the cohesion between credit card customers and the bank, has become a topic of great interest [10]. Shaw [8] mainly describes how to incorporate data mining into the framework of marketing knowledge management.

The collective application of association rule techniques reinforces the knowledge management process and allows marketing personnel to know their customers well to provide better quality services. Song [6] proposed a method to illustrate change of customer behavior at different time snapshots from customer profiles and sales data.
The basic idea is to discover changes from two datasets and generate rules from each dataset to carry out rule matching.

4. CONCLUSIONS AND FURTHER WORK

In this work a revision on the main application areas of association rules has been focused. It is all about to find some kind of pattern or relationship between various datasets. The outcome is association rules, and it is an iterative refinement process.

Further work can be done on the employee database for finding association rules related to job stability.

REFERENCES

[1] C. Branden and J. Tooze, “Introduction to Protein Structure”, Garland Publishing inc, New York and London, 1991.

[2] D. Gamberger, N. Lavrac, and V. Jovanoski, “High confidence association rules for medical diagnosis”, *In Proceedings of IDAMAP99*, pages 42-51.

[3] D. Malerba, F. Esposito and F.A. Lisi, “Mining spatial association rules in census data”, *In Proceedings of Joint Conf. on "New Techniques and Technologies for Statistcs and Exchange of Technology and Know-how”*, 2001.

[4] G. Saporta, “Data mining and official statistics”, *In Proceedings of Quinta Conferenza Nationale di Statistica,* ISTAT, Roma, 15 Nov. 2000.

[5] G. Serban, I. G. Czibula, and A. Campan, "A Programming Interface For Medical diagnosis Prediction", Studia Universitatis, "Babes-Bolyai", Informatica, LI(1), pages 21-30, 2006.

[6] H. S. Song, J. K. Kim and S. H. Kim, “Mining the change of customer behavior in an internet shopping mall”, Expert Systems with Applications, 2001.

[7] http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm

[8] M. J. Shaw, C. Subramaniam , G. W. Tan and M. E. Welge, “Knowledge management and data mining for marketing”, Decision Support Systems, v.31 n.1, pages 127-137, 2001.

[9] N. Gupta, N. Mangal, K. Tiwari and P. Mitra, “Mining Quantitative Association Rules in Protein Sequences”, *In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining – AUSDM,* 2006

[10] R. S. Chen, R. C. Wu and J. Y. Chen, “Data Mining Application in Customer Relationship Management Of Credit Card Business”, *In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, Volume 2, pages 39-40.