Use MADM method to Decision making among results extracted from Association rules

Hiva Samadian
Tarbiat Modares University, Tehran, Iran
H.Samadian@modares.ac.ir

Hamid Abed
Tarbiat Modares University, Tehran, Iran
H.Abed@modares.ac.ir

ABSTRACT

In the arena of data mining, ranking and choosing the rules are of essential importance. For ranking the association rules, traditionally, two criteria of confidence and support have been take use of, but in some functional arenas, it is necessary to involve other criteria through the process of ranking. As the number of criteria increases drawing on multi attribute decision making methods would be compelling. In the present paper, a ranking approach based on AHP, as a defensible method for measuring the efficiency of association rules for multi attribute criteria, will be presented.

Keywords- Association rules, multi attribute decision making, AHP, data mining

1. INTRODUCTION

Thanks to the technology advancements, automata measurement and information storing have become possible in different arenas of science and hence the human is drawn in a world of information of different scientific fields and the ways have been paved for more accurate modeling of complex systems, based on information. Data mining, in contrast with structured surveys of data, has this distinction that through data mining, rather than questioning a certain relation, the valuable patterns and connections between the relations, which are not predetermined, will be extracted (Agrawel and Imielinski, 1993). Data mining could be explained as a type of automata inductive learning.

Governmental agencies, scientific institutions and enterprises, all have their own huge private sources of information for collecting and storing the data. But in practice, a tiny part of these data is taken use of because in many cases, due to the large volume and complex structure of information, an efficient management and analysis on the data is not possible or would be so difficult. In today's new competitive world the potential for extracting the knowledge laying behind these data could not be denied, neither working on them could be overlooked. The overall process for employing an automata approach, which involves new methods of knowledge extraction from the data, is called data mining. Data mining is a repetitive process in which in every step a discovery is made by manual or automata procedures. Data mining is especially useful in explorational analysis where there is no predefined

Assumption concerning the outputs and results. Data mining, by definition, is searching for valuable and non-evident information on large sets and sources of data. This is a joint activity between human and machine (computer). By a proper combination of sophisticated people's knowledge in defining the problem and the goals, and the searching power of computers, we can attain the best result.

Today, data mining techniques have found vast applications in businesses. Using data mining techniques, the knowledge is extracted in sets of rules. However, due to budget and resources limitations, only few of these rules are utilized as a guide for implementation practices. Since extraction of association rules, by providing necessary data and information, makes choosing possible for decision makers, it is considered different from automata learning actions (Webb and Zhang, 2005).

2. ASSOCIATION RULES

Association Rules Mining is one of the most useful and popular methods in data mining. The association rules were first employed for discovering the costumers shopping patterns, using the analysis of stores sales data.

It was observed that Frequent Itemset plays a critical role in many of the data mining methods, some of them which their main effort was focused on discovering interesting patterns from databases. Among these methods, methods of Association Rules, Correlation, Sequences, Episodes and ... could be mentioned.

In fact, association rules express patterns of how products are being shopped together. For instance the association rule "(80%) cake→ drink" indicates that among 5 drink shoppers, 4 will buy cake too. Such rules could have wide applications, which among them all the improvement of decision quality could be mentioned.

In general, the extraction of association rules leads to the discovery of a set of rules that grant a specific set of criteria and limitations, such as the least support and the least confidence, but is not capable of producing a specific model appropriate for the data and the business environment.

Today, the attractiveness and efficiency assessment of association rules in the field of data mining, due to its goal of refining and creating the invaluable data, is of great importance. In many trading and business applications in real world, the ranking of rules resulting from data mining, based on a set of qualitative criteria (Tan and Kumar, 2000) and limitations of trading resources (Choi and Kim, 2005), is essential. Choosing more valuable rules for application and execution increases the possibility of success in data mining projects.

Since products with high profit and added value are infrequently sold, the association rules extracted with limitations of the least support and confidence will not be the rules of profitability. (Cohen et al. 2000) describes the problem of shopping drinks and caviar as a proof to this issue. Shoppers infrequently buy these two in a same place. While these two are not among the products with mediate abundance of shopping, their profit is much more than products with lower value and higher abundance. In fact, shopping carts with low abundance, due to the functional issues such as the profit, sometimes are more popular (Webb and Zhang, 2005), and this is while the traditional algorithms for association rules in data mining, without considering the knowledge of the problem domain, cannot categorize such thresholds for products in the form of favorable unity sets (association products). If we decrease the threshold for the sensitivity of limitations, we will confront a multitude of rules, which makes the analysis too complicated.

In most of the available algorithms used for extracting the association rules, high support criterion will be utilized for choosing sets of elements with high frequency, and all the elements in a set are of identical weight (Tao et al. 2003). Thus, former approaches are not suitable for extracting important, but with low frequency, rules.

This paper is authored with the aim of presenting a ranking model using the technique Hierarchy Analysis Process. In this model after extracting the rules using the algorithms of association rules, the process of hierarchy analysis, upon specific criteria, will be employed for ranking the rules.

3. FORMULATING THE PROBLEM

In every research, in order to find a solution for the selected problem, we need to define and formulate the problem, so that via these formulas we would be able to test the theories presented as possible and temporary solutions for the problem. For this reason, the definitions and symbols used in this paper, along with the method of formulation, is described as follows:

Definition 1: Assume $I = \{i_1, i_2, ..., i_n\}$ as a set of items with n members, the set $X = \{x_1, x_2, ..., x_k\} \subseteq I$ is called an itemset or a set with k items.

Definition 2: A transaction on I is a pair T= (tid, I) in which tid is the number of transactions and I is an itemset on I. If item $X \subseteq I$ is a member of I, then $X \subseteq I$ supports transaction T from the itemset X.

Definition 3: The cover of Itemset in database D is defined as below:

$$Cover(X, D) = \{tid \,|\, (tid, I) \in D, X \subseteq I\}$$

Definition 4: Supporting an itemset in D means the number of transactions that cover X in D.

$$Support(X, D) = |Cover(X, D)|$$

Definition 5: The abundance means the possibility of occurrence of X in a T transaction belonging to D:

$$frequency(X, D) = p(x) = \frac{Support(X, D)}{|D|}$$

In which |D| is the total number of transactions and equals |D|=support ({}, D).

Definition 6: An itemset is said to be repetitive if its amount of support is no less than a threshold amount of σ_{abs}, where $0 < \sigma_{abs} <$ |D|.

When speaking of abundance, this amount could be expressed as $0 \le \sigma_{rel} \le 1$ and naturally the relation $\sigma_{abs} = [\sigma_{rel} \times |D|]$ will be true.

Definition 7: Assume D to be a transaction bank on items I and σ is a threshold amount, then the set of repetitive items will be shown as below:

$$F(D, \sigma) = \{X \subseteq I \,|\, Support(X, D) \ge \sigma\}$$

Sometimes, F (D, σ) is shown with F also.

Example 1- (data set mining)

If I, D and σ are given, find F (D, σ).

In practice, F (D, σ) is not the only important factor, but the real number of frequency of items present in F (D, σ) is more important. X =>Y is an association rule in which X and Y are itemsets and $X \cap Y = \emptyset$. This rule expresses that if the transaction contains X it should contain Y too. The amount of support for rule X =>Y equals to the amount of support $X \cap Y$ in D and its frequency is equal to the frequency of $X \cap Y$ in D.

Definition 8: The confidence coefficient or the accuracy of a rule X=>Y in D means the possibility of presence of Y in a transaction in case of presence of X in that transaction.

Definition 9: A rule is said to be confident if P (Y|X) is more than an amount of threshold, say λ where $0 \le \lambda \le 1$.

Definition 10: Assume D as a transaction bank on I, σ as a support threshold and λ as a confidence threshold, the set of repetitive and confident rules, regarding σ and λ will be as below:

Example 2- (association rules mining)

Assume that a set I, the set of D transactions on it, σ and λ are give, find $R(D, \sigma, \lambda)$.

In addition to R, the support and confidence coefficient of extracted rules are also important. As you see the issue of repetitive itemset mining is a particular case of association rules mining issue, meaning that each repetitive item represents a rule, even unvalued, as X =>{}, where the confidence coefficient for this rule is 100%. It is evident that the number of frequency of the rule and of X is identical.

It is also obvious that for every repetitive itemset I, there are rules as X =>Y where $X \cup Y$=I and they all have a confidence coefficient of at least σ_{rel}.

4. The AHP Method

The AHP evaluation method was introduced and founded by Saati in 1980. The basic of this method is performing pair comparisons and determining the preferability of elements to each other in reference to the criteria under question. This method is taken use of for assessment problems and problems concerning the preferability of multiple choices regarding the criteria under question, which its self could contain other sub-criteria. Firstly, we proceed with a revision on the basic steps in AHP. The basic steps in AHP are as below:

Step 1: determining the goals, criteria, sub-criteria and the choices, from the available information of the problem.

Step 2: creating the hierarchical graphical show for the problem.

The hierarchical graphical show is a simple presentation from the real complex problem, in which in the head the overall goal of the problem and in next levels the criteria and sub-criteria, and in the bottom the choices, are placed.

Step 3: performing the pair comparisons.

For pair comparison of elements in each level Saati proposed the following measure:

Through the pair comparison of the elements, if we comparison ith element with jth element, one of the below conditions could describe the significance (preferability) of element i to element j:

1. There is extremely preferability
2. There is a very strong preferability
3. There is a strong preferability
4. There is a little preferability
5. There is an equal preferability

For the above judgments Saati used quantitative values respectively as follow: 1, 3, 5, 7, 9, i.e. as for the amount of ith element preferability to jth element, one of the numbers 1, 3, 5, 7, 9 could express the amount of preferability. Therefore, as for the preferability of element j to element i, according to the converse principle in AHP, respectively the numbers 1, 1.3, 1.5, 1.7 and 1.9 will be used and the numbers 2, 4, 6 and 8 will be used as medial values for the comparison. For instance, if element i have a strong preferability to element j then number 5 would be assigned. Therefore, the elements pair comparison matrix for each level would be made this way.

Step 4: determining the weights:

Calculating the weights for the process of hierarchy analysis involves two parts of relative weight calculation and absolute (final) weight calculation. The relative weights would be obtained from the pair comparison matrix, while the absolute weight is the final rank of each choice, which is obtained from the compilation of relative weights. The final weight of each choice in a hierarchy analysis is obtained from the addition of the sum of the criteria significance and the choices weight. There are various methods for calculating the relative weights, which among them all the below methods are mentioned:

1- Least Squares Method
2- Logarithmic Least Squares Method
3- Eigenvalue Method
4- Approximation Method

The first three methods produce weights accurately, approximation methods are among former methods, nevertheless due to the ease and much lower volume of calculations, these methods, containing methods geometric mean and simple mean are still used

5. THE PROPOSED APPROACH

The attractiveness of a rule could be utilized for filtering a high number of rules and presenting the important ones for the decision maker (Mitra and Paul 2003). The two thresholds of support and confidence are chosen only from two the viewpoints of database science. Thus, the information from the problem domain could provide more functional criteria for ranking the association rules. The method of association rules extraction, which is executed by the Apriory algorithm, in theory, except the two mentioned thresholds, does not present a framework or methodology for choosing the rules. In this paper, combining the association rules extraction technique and technique AHP, a propositional approach is presented. This general approach is depicted in figure 1 and the steps are as below:

1- Entering and preparing the data for rules extraction
2- Association rules extraction using the Apriory algorithm using the least support and confidence
3- Determining the thematic attractiveness criteria with regard the knowledge and problem information and the opinion of decision makers
4- Performing technique AHP with regard to the rules cluster as the cluster for choices and criteria
5- Calculating the score from the super matrix
6- Choosing and ranking the rules for implementation, upon the obtained scores in step 5.

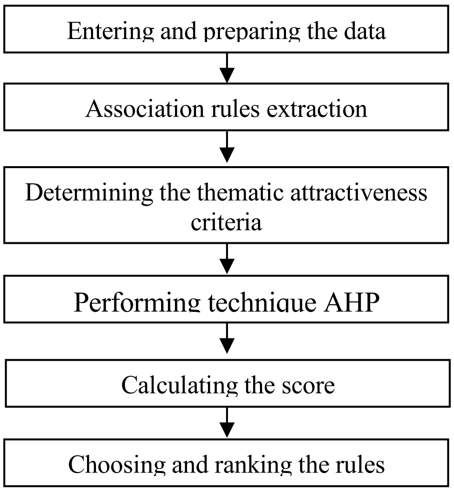
Entering and preparing the data
Association rules extraction
Determining the thematic attractiveness criteria
Performing technique AHP
Calculating the score
Choosing and ranking the rules

Fig.1. The proposed approach steps

In fact, the attractiveness of rules, which is beyond the discussion of rules frequency, is measured according to the criteria specified in steps 4 and 5, while in the past, for extracting the rules, only the Apriory algorithm was taken use of and this approach was not considered with a methodology, and for this very reason the rules never had low abundance and the problem described before (the shopping chart with low abundance but high profit), still remained a problem.

In shopping chart analysis, understanding which products are sold together and which products are of profit for the seller, are both attractiveness criteria of the analyzer. Since the purpose of this paper is presenting the shopping chart analysis approach, the below information, regarding the problem domain, are proposed and their structure is depicted in figure 2:

1- Trading and monetary value describes the profitability of a rule.
    - The criteria of expecting monetary value- with an assumption of independence, the expectational profit for shopping product X equals to the probability of shopping Y with the condition of X being shopped multiplied by the profit of product Y.
    - The monetary value criterion of an increase- the profit of one rule

minus by the regular profit of a costumer.

2- The frequency- implies on the statistical index of the occurrence of a rule
   - The support criterion
   - The confidence criterion
   - The rate of interest criterion-this index indicates the ratio of rules common probability to their probability of independence occurrence.

3- Recency- the word Recency implies on the concept of the time of taking use of on rule at the recent time range.
   - The degree of change criterion- in data mining, seldom, the emphasize is on the occurrence of a rule at the recent times, while only the total abundance is seen, this is while the criterion of the degree of change points to the Recency of usage and the growing trend of a rule in the future (Choi et al. 2005).

4- Monetary value of increase criterion - this criterion is expressed as the profit of one rule minus by the regular profit of a costumer.

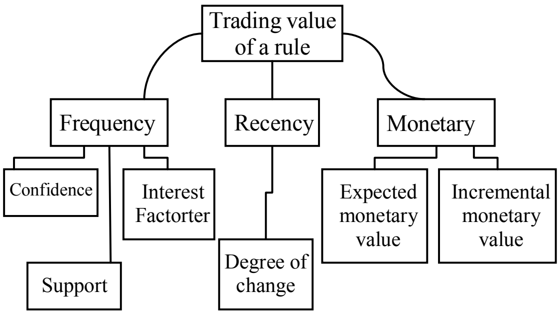

Fig.2. The criteria structure in the proposed approach

6. CONCLUSION

Extraction of association rules is one of the popular techniques recently developed in the arena of data mining. Attractiveness assessment and the functionality of association rules are among the important issues in data mining applications. Through the analysis of shopping chart, marketing and sales analyzers are not satisfied with only a set of pattern rules extracted from algorithms, but they rather prefer to specify the functionality of the rules, through the domain of their problem, using the ranking of the rules and based on specific criteria. This is while; the complexity of rules and criteria makes this work too difficult without taking use of a specific methodology or approach. Former approaches ignored the knowledge of problem domain for choosing the rules, and by relying on only two statistical criteria of confidence and support; they proceeded to choose the rules. In this paper, in order to fill this gap, a propositional approach is developed, which by taking advantage of technique AHP for multi attribute decision making, and also the criteria structure for the problem of shopping

Chart, could play as a proper instrument for attractiveness and functionality assessment of association rules. The proposed approach provides a good base for extracted rules and could be efficient through their evaluation and selection.

References

[1] Agrawel, R., Imielinski, T., & Swami, A; “Mining association rules between sets of items in large databases”, Proceedings of the ACM SIGMOD Conference on Management of data, 254-259, 1993.

[2] Chen, M.-C., & Wu, H.-P; “An association-based chustering approach to order batching considering customer emand patterns”, Omaga-International Journal of Management Science, 33(4), 333-343, 2005.

[3] Choi, D. H., Ahn, B. S., & Kim, S. H.; “Prioritization of association rules in data mining: multiple criteria decision approach” Expert System with Applications, 29(4). 876-878, 2005.

[4] Cohen, E., Datar, M., Fujiwara, S., Gionios, A., Indyk, R., Motwani, P., Ullman, J., & Yang, C.; “Finding interesting associations without support pruning”. In Proceeding of the 16th international conference on data engineering, pp. 489-500, 2000.

[5] Mitra, E., Paul S.; “On supporting interactive association rule mining. Lecture Notes in Computer science”, 1874, 307-319, 2003.

[6] Tao, F., Murtagh, F., & Farid, M.; “Weighted association rule mining using weighted support and significance framework”. In Proceeding of the ACM SGMOD international conference on management of data, Simod-03, pp. 661-666, 2003.

[7] Tan, P. N., & Kumar, V.; “Interestingness measures for association patterns: A perspective”, KDD 2000 workshop on post processing in machine learning and mining, Boston, MA, 2000.

[8] Webb, G. I., Zhang, S.; “K-optimal rule discovery. Data Mining and Knowledge Discovery”, 10(1), 39-79, 2005.