

# Analysis of the Performance of Various Algorithms and Interestingness Measures in Association Rule Mining

<sup>1</sup>Mukesh Sharma, <sup>2</sup>Jyoti Choudhary, <sup>3</sup>Gunjan Sharma

<sup>1,2,3</sup>Dept. of CSE, The Technological Institute of Textile & Sciences, Bhiwani Maharshi Dayanand University, Rohtak, Haryana, India

## Abstract

Association rule mining is one of the most important and well researched techniques of data mining and was first introduced by Agrawal in 1993. It aims to find out interesting correlations, frequent pattern, casual structures or associations among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as market and risk management, telecommunication networks, inventory control and weather forecasting etc. So, it becomes important to choose the best algorithm to find the interesting rules. This paper discusses the various parameters for measuring the interestingness of association rules and also the various association algorithms.

## Keywords

Data Mining, Association Rule Mining, Correlation, Frequent Pattern, Interestingness Measurement

## I. Introduction

Data mining is defined as the process of extracting hidden predictive information from large databases and this powerful technique helps companies to gather the most important information in their data warehouses [7]. There exist various data mining tools which help in predicting future trends and behaviors, allowing businesses to make knowledge-driven decisions. Data mining tools can solve business problems that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that business experts may miss because it lies outside their expectations. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and Knowledge Discovery in Databases (or KDD) are considered similar, data mining is actually part of the knowledge discovery process.

The Knowledge Discovery in Databases process consists of a few steps starting from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

### A. Data Cleaning

At this step, noise data and irrelevant data are removed from the huge database.

### B. Data Integration

At this step, data from multiple sources, often heterogeneous, may be combined in a common source.

### C. Data Selection

At this step, the data relevant to the analysis is selected and retrieved from the data collection.

### D. Data Transformation

Is also known as data consolidation, it is a phase in which the selected data is transformed into forms suitable for the mining

procedure.

### E. Data Mining

It is the crucial step in which techniques are applied to extract potentially useful patterns.

### F. Pattern Evaluation

At this step, strongly interesting patterns representing knowledge are identified based on given measures.

### G. Knowledge Representation

This is the final step in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the results of data mining.

Data mining techniques can be implemented rapidly on existing hardware and software platforms to enhance the value of existing information resources, and can be integrated with new products and systems. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze huge databases to deliver answers to questions such as, "Which clients are most likely to respond to the next promotional mailing, and why?"

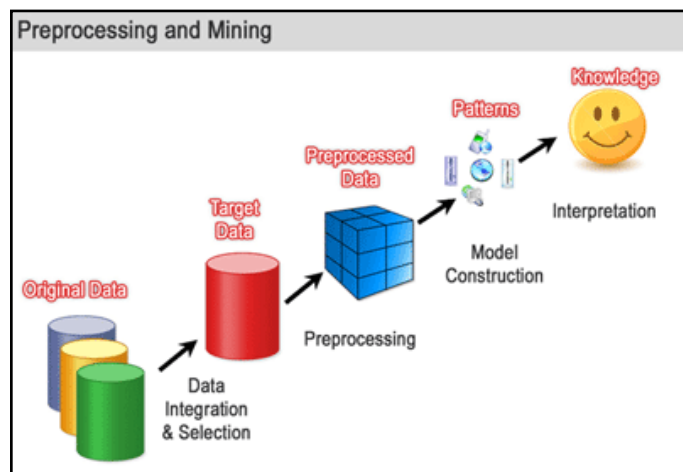


Fig. 1:

## II. Association Rule Mining

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on the threshold value called support, identifies the frequent item sets [8]. There also exists another threshold value called confidence, which is the conditional probability that an item appears in a transaction when another item appears, is used to find out association rules. Association analysis is mainly used for market basket analysis. For example, it could be useful for the Audio/Video Store manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form:

$X \rightarrow Y [s, c]$ , where X and Y are conjunctions of attribute value-pairs, and s (for support) is the probability that X and Y appear together in a transaction and c (for confidence) is the conditional probability that Y appears in a transaction when X is present.

### III. Various Interestingness Measures in Association Rule Mining

#### A. Support

Support for ARM is introduced by Agrawal in 1993 [1] and is defined as the proportion of transactions in the data set which contain the itemset.. It measures the frequency of association, i.e. how many times the particular item has been occurred in a dataset. An itemset with large support is called frequent itemset. In terms of probability theory, it can be defined as:

Support =  $P(A \cap B)$  = number of transactions containing both A and B / Total number of transactions.

#### B. Confidence

Confidence basically measures the strength of the association rules. It is defined as the fraction of the transactions that include both A and B to the total number of records that contain A. It determines how frequently item B occurs in the transaction that contains A. Confidence expresses the conditional probability of an item. The definition of confidence is

$$\text{Confidence} = P(A | B) = \frac{P(A \cap B)}{P(B)}$$

#### C. Predictive Accuracy

Predictive accuracy is also another way to measure interestingness of an association rule. Basically this accuracy is used for the Predictive Apriori Algorithm. According to Scheffer, the definition of predictive accuracy is:

Let D be a data file with n number of records. If  $[a \rightarrow b]$  is an Association Rule which is generated by a static process P then the predictive accuracy of  $[a \rightarrow b]$  is  $c([a \rightarrow b]) = \frac{P_n[n \text{ satisfies } b | n \text{ satisfies } a]}{P_n[n \text{ satisfies } a]}$  where distribution of n is governed by the static process P and the Predictive Accuracy is the conditional probability of  $a \rightarrow n$  and  $b \rightarrow n$ .

#### D. Lift

The lift value basically defines the importance of a rule. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the rule body and the rule head divided by the support of the rule body.

$$\text{lift} = \frac{\text{confidence}}{\text{expected\_confidence}} = \frac{\text{confidence}}{(s(\text{body}) * s(\text{head}) / s(\text{body}))} = \frac{\text{confidence}}{s(\text{head})}$$

$s(\text{body})$  :- is the support of the rule body

$s(\text{head})$ :- is the support of the rule head

The expected confidence is identical to the support of the rule head. It is assumed in the definition of the expected confidence that there is no statistical relation between the rule body and the rule head. This means that the occurrence of the rule body does not influence the probability for the occurrence of the rule head and vice versa. The lift is a measure for the deviation of the rule from the model of statistical independency of the rule body and rule head. The lift is a value between 0 and infinity:

- A lift value greater than 1 indicates that the rule body and the rule head appear more often together than expected, this means that the occurrence of the rule body has a positive

effect on the occurrence of the rule head.

- A lift smaller than 1 indicates that the rule body and the rule head appear less often together than expected, this means that the occurrence of the rule body has a negative effect on the occurrence of the rule head.
- A lift value near 1 indicates that the rule body and the rule head appear almost as often together as expected, this means that the occurrence of the rule body has almost no effect on the occurrence of the rule head.

### IV. Association Algorithms

#### A. Apriori Association Rule

Apriori is an association rule algorithm proposed by R. Agrawal and R. Srikant in 1993 [1] for mining frequent item sets for boolean association rule. The name of algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori employs an iterative approach known as level-wise search, where k item set are used to explore (k+1) item sets. There are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step the occurrence of each candidate set in database is counted and then pruning of all disqualified candidates (i.e. all infrequent item sets) is done. Apriori uses two pruning techniques, first on the basis of support count (should be greater than user specified support threshold) and second for an item set to be frequent, all its subset should be in last frequent item set. The iterations begin with size 2 item sets and the size is incremented after each iteration. This algorithm is easy to implement and parallelized but it has the weakness that it requires various scans of databases and is memory resident.

#### B. Predictive Apriori Association Rule

In predictive Apriori association rule algorithm, support & confidence both are combined into a single measure called predictive accuracy. This predictive accuracy is used to generate the Apriori association rule. In Weka, this algorithm generates 'n' best association rule where n is selected by the user [9].

#### C. Filtered Associator

This algorithm is used for running an arbitrary associator on data that has been passed through an arbitrary filter. Like the associator, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure [9]. Here in this algorithm, the Apriori, Predictive Apriori & Tertius association rule algorithm can be used for getting the result.

#### D. Tertius Algorithm

Tertius is basically a first order logic discovery algorithm [6]. Tertius employs a complete top-down A\* search over the space of possible rules [4]. The Tertius algorithm builds rules out of the attribute pair values in the training data and ranks them according to their reliability, that is how many times the rule holds true in the training data.

A rule consists of two parts a body and a head. The body contains the conditions (which are known as literals) required for the rule to hold, and can consist of any number of literals. The head contains the event that occurs when the rules hold true. During rule learning, Tertius starts with an empty rule – means a blank body and a blank head. The rule is then refined by adding attribute-value pairs in the order that they appear in the dataset. Once this completes, the algorithm counts the number of times the rule holds true (both

body and head are true) and the times when the rule gives a false positive (when the body is true but the head is false).

### 1. Disadvantage

One disadvantage of Tertius is its relatively long runtime, which is mainly dependent on the number of literals in the rules. Increasing the number of literals increases the runtime exponentially. Tertius can take several hours for larger datasets.



Gunjan Sharma received her B.Tech degree from the Technological Institute of Textile & Sciences, Bhiwani, India. Now She is pursuing her M.tech degree from the Technological Institute of Textile & Sciences, Bhiwani, India. Her research interests include various data mining techniques, association rule mining and statistical techniques on datasets.

### VI. Conclusion

Association rule mining is really the emergeable topic now a days. Researchers aim to find the best and strong association rules. Association analysis can generate large quantity of rules, most of which are of no interest to the user. So it is required that strong rules should be find out. Interestingness measures are used to find the truly interesting rules. This paper presents a review of the various interestingness measures and also discusses the various association algorithms and their advantages and weaknesses.

### References

- [1] Agrawal, R., Imielinski, T., and Swami, A. N., "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993.
- [2] Agrawal, R., Srikant, R., "Fast algorithms for mining association rules", In Proc. 20th Int. Conf. Very Large Data Bases, 487-499, 1994. Based Systems 12(5-6), 309-315 (1999).
- [3] Agrawal, R., Imielinski, T., Swami, A. N., "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993.
- [4] Li Yang, Mustafa Sanver, "Mining Short Association Rules with One Database Scan", Int'l Conf. on Information and Knowledge Engineering; June 2004.
- [5] Goswami D.N, Chaturvedi Anshu, "An Algorithm for Frequent Pattern Mining Based On Apriori", (IJCSSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947
- [6] P.A. Flach, N.Lachiche, "confirmation-guided discovery of first-order rules with tertius", Kluwer Academic Publishers. The Netherlands, Vol. 42, pp. 61-95, 2001.
- [7] Margaret H. Dunham "Data Mining Introductory and Advanced Topics".
- [8] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition "Elsevier publications", 2006.
- [9] Weka, [Online] Available: <http://www.cs.waikato.ac.nz/ml/weka/>.
- [10] [Online] Available: [http://www.cs.sunysb.edu/~cse634/lecture\\_notes/07apriori.pdf](http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf) accessed on date 02-02-2012