ISSN-2230-7540

Human Involvement in Association Rule Mining

Luxmi Parmar
Research Scholar, CMJ University, Shillong, Meghalaya

Dr. Rajesh Pathak
Research Scholar, CMJ University, Shillong, Meghalaya

INTRODUCTION

A number of data mining algorithms have been introduced to the community that perform summarisation and classification of data with respect to a target attribute, deviation detection, and other forms of data characterisation and interpretation. One popular summarisation and pattern extraction algorithm is the association rule algorithm. Association rule is described as an associational relationship between a group of objects in a transactional database [Zhang S. et.al., 2006]. The following discussion shall be useful to Justify the intervention of human heuristics in the data mining process.

OBJECTIVE

Study the features of the data mining tools: TANAGRA and MATLAB – SOM TOOLBOX and FUZZY LOGIC TOOLBOX; and find the ponits where interaction is needed.

Develop and execute interactive data mining algorithms in "*java*" paltform and report the results obtained.

Apply and analyse interactive data mining process to various medical databases arranged from different sources grasp the basic idea behind association rule mining algorithm.

MATERIAL AND METHOD

To perform the experimentation for the present research work, various data mining techniques as well as tools were well thought-out. Different data mining techniques such as association rule mining, clustering was implemented in programming language (Java Platform). The data mining algorithms with human interaction points were designed and tested on various databases.

The data mining market consists of software vendors offering tools that extract predictive information from large data stores, which can then be analysed to enhance corporate data resources and generate predictions regarding business trends and behaviour. Specifically, these tools provide statistical data models (classification or clustering studies, linear regression, and current or predictive modeling) and utilise visualisation functions to support the analysis of massive quantities of data stored by business organisations. Data mining tools may be implemented on existing customer platforms or integrated with other applications as part of a larger data quality initiative or business intelligence (BI) strategy. Data mining tools provide both developers and business users with an interface for discovering, manipulating, and analysing corporate data.

Although there are a number of data mining tools available in the market, but the following tools were used to perform the experimentation of this research work (which suits the problem best),and tested on points of human interactivity

Let D be a transaction database and I = $\{i_1, i_2, ..i_m\}$ be an item set. Transaction database D contains a sequence of transactions T = $\{t_1, t_2, .. t_n\}$ (where T $\subseteq$ I) with a sole identifier. An association rule X→Y may be discovered in the data where X and Y are conjunctions of items and

$X \cap Y = \Phi$. The intuitive meaning of such a rule is that transactions in the database which contains the items in X tend to also contain the items in Y. The user supplies minimum support and confidence thresholds. The support of the rule X→Y represents the percentage of transactions from the original database that contain both X and Y. The confidence of the rule X→Y represents the percentage of transactions containing items in X that also contains items in Y. Association rules are based upon the concept of strong rule. A rule that satisfies both minimum support and minimum confidence at the same time has been described as a strong rule in the literature [Agrawal R. et.al., 1993].

The process of discovering of association rules is broken up into two steps [Agrawal R. et.al., 1994]:

(i) Find all itemsets (set of items appearing together in a transaction) whose support is greater than the specified threshold. Itemsets with minimum support are called *frequent item sets*.

(ii) Generate association rule from the frequent item sets. To do this, consider all partitioning of the item sets into left-hand and right-hand sides. The confidence of a rule X→Y that satisfies minimum support is calculated as support (XY)/support (Y). All the rules that meet the confidence threshold are reported as discoveries of the algorithm.

Association rules were first introduced in [Agrawal R. et.al., 1993]. The subsequent paper [Agrawal R. et.al., 1994] discusses Apriori algorithm that is considered as one of the most important contributions to the field of data mining. Although, other algorithms such as AIS [Agrawal R. et.al., 1993] and SETM [Houtsma M.A.W. et.al., 1993] are also available for mining association rules, yet Apriori remains the most widely used approach for generating frequent itemsets. The algorithm accomplishes the search of frequent itemsets in recursive order. It first scans the database D and calculates the support of each single item in every record I in D, and denotes it as C_1. Out of the itemsets in C_1, the algorithm computes the set L_1 containing the frequent 1-itemsets. In the k^{th} scan of the database, it generates all the new itemset candidates using the set L_{k-1} of frequent (k-1) itemsets discovered in the previous scanning and denotes it as

C_k. And the itemsets whose support is greater than the minimum support threshold are kept in L_k. This process is repeated until no new frequent itemsets are found.

Table 1: Notations used in Apriori algorithm.

Name	Description
k-itemset	An itemset having k items.
L_k	Set of large k-itemsets (those with minimum support). Each member of this set has two fields: i) itemset and ii) support count.
C_k	Set of candidate k-itemsets (potentially large itemsets). Each member of this set has two fields: i) itemset and ii) support count.
U_k L_k	Set of generated itemsets.

Table 2: Example dataset D.

T_{id}	Items
10	AB
20	ABE
30	ABCE
40	CD

The Apriori approach of searching frequent itemsets is explained with the database of Table 2. The algorithm assumes the minimum support threshold to be "*2*". Firstly, it initialises C_1 as the set of all items, takes count of elements in it, and puts in L_1 the elements satisfying the minimum support. Thereafter, set C_2 is generated using L_1 and count of the elements is computed from the

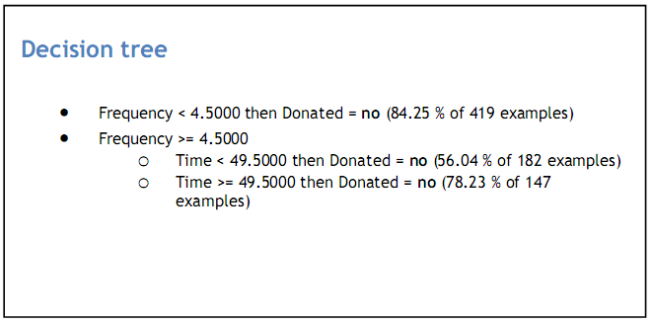
ISSN-2230-7540

scan of database *D*. The frequent itemsets from C_2 are kept in the set L_2. In the similar way, L_3 is generated. As there is a single itemset in L_3, the set C_4 is empty. So, this arithmetic comes to an end (min_support = 2). Working of the algorithm has been explained in Figure 1.

CONCLUSION

In this experiment, ID3 decision tree algorithm has been run using TANAGRA data mining software. In this application domain, the purpose was to find the input factors which determine the blood donation of a person. Therefore, '*Donated*' attribute was set as the target attribute, and 'F*requency*' and '*Time*' were set as input to the algorithm. The Figure 3 shows the results of the ID3 decision tree algorithm on the training dataset (randomly selected sample from the database). By observing the results, one can see that if '*Frequency*' is less than 4.5000, then, 84.25% chances are that the person will not donate the blood in the current month.

Figure 3: Decision tree for the user-specified attributes.



The knowledge derived this way can be used to form rules in the whole database. Such rules can help the domain users to predict the future cases in the same domain. The following rule i.e. *Rule2* can be formed from the decision tree constructed (Figure 3).