International Journal of Soft Computing 11 (3): 155-159, 2016
ISSN: 1816-9503

A New Approach Towards Item Set Mining Using Distribution Model

Paul P. Mathai and R.V. Sivabalan
Department of Computer Science and Engineering,
Noorul Islam University, Kanyakumari, Tamil Nadu India

Abstract: Data mining can be defined as an activity that extracts some new nontrivial information contained in large databases. Traditional data mining techniques have focused largely on detecting the statistical correlations between the items that are more frequent in the transaction databases. Generally, several applications are using data mining in different fields like medical, marketing and so on. Numerous methods and techniques have been developed for mining the information from the databases. In this study, we propose a new approach for itemset mining on utility and frequency using distribution model and association rule mining based research works.

Key words: Data mining, Knowledge Discovery Database (KDD), itemsets, utility, frequency

INTRODUCTION

In the business world, corporate and customer data are becoming recognized as a trategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer based methodology, including new techniques for discovering knowledge from data is called data mining (Khalilzadeh and Fard, 2008). The objective of data mining is to identify valid novel, potentially useful and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology and data pattern processing). The term “data mining ” is primarily used by statisticians,database researchers and the MIS and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data where data mining is a particular step in this process (Jackson, 2002; Fayyad *et al*., 1996). Data mining is a highly inter disciplinary area spanning a range of disciplines; statistics, machine learning, databases, pattern recognition and other areas (Deogun *et al*., 1997).

In general, data mining methods can be classified into two categories; predictive and descriptive. Predictive data mining methods predicts the values of data, using some already known results that have been found using a different set of data. Predictive data mining tasks include: classification, prediction. Descriptive mining tasks characterize the general properties of the data in database. This is done by identifying the patterns and relationships in the data (Velickov and Solomatine, 2000). In data mining, items are mined form the database based on two constraints: items frequency and utility.

Calculating itemset support (or frequency counting) is a fundamental operation that directly impacts space and time requirements of many widely used data mining algorithms. Some data mining algorithms (i.e., frequent itemset mining) are only concerned with identifying the support of a given query itemset while others (i.e., pattern-based clustering algorithms must in additionidentify the transactions that contain the query itemset (Malik and Kender, 2007). The goal of frequent itemset mining is to find items that co-occur in a transaction database above a user given frequency threshold without considering the quantity or weight such as profit of the items. However, quantity and weight are significant for addressing real world decision problems that require maximizing the utility in an organization. The high utility itemset mining problem is to find all itemsets that have utility larger than a user specified value of minimum utility (Erwin *et al*., 2007, 2008).

Utility-based data mining is a broad topic that covers all aspects of economic utility in data mining. It encompasses predictive and descriptive methods for data mining, among the later especially detection of rare events of high utility (e.g., high utility patterns) (Podpecan *et al*., 2007). Utility based data mining refers to allowing a user to conveniently express his or her perspectives

Corresponding Author: Paul P. Mathai, Department of Computer Science and Engineering, Noorul Islam University, Kanyakumari, Tamil Nadu, India

concerning the usefulness of patterns as utility values and then finding patterns with utility values higher than a threshold. A pattern is of utility to a person if its use by that person contributes to reaching a goal (Yao *et al*., 2006).

Mining based on itemsets frequency: In data mining regular itemset mining is a conventional and significant problem. An itemset is repeated if its support is not less than a brink stated by users. Conventional regular itemset mining approaches have chiefly regarded as the crisis of mining static operation databases. In the operation data set regular itemsets are the itemsets that happen often. To recognize all the regular itemsets in a operation dataset is the objective of frequent itemset mining. Within the finding of relationship rules it created as a phase but has been simplified autonomous of these to several other samples. It is confronting to enlarge scalable methods for mining regular itemsets in a huge operation database as there are frequently a great number of diverse single items in a distinctive transaction database and their groupings may form a very vast number of itemsets.

Mining based on itemsets utility: Depending upon his circumstance of usage as precised by the user a high utility itemset is the one with utility value larger than the minimum brink utility. A wide topic that wraps all features of economic utility in data mining is known to be utility-based data mining. It includes the work in cost-sensitive education and dynamic learning as well as work on the recognition of uncommon events of high effectiveness value by itself. By maintaining this in mind, we at this point offer a set of algorithms for mining all sorts of utility and frequency based itemsets from a trade business deal database which would considerably aid in inventory control and sales promotion. Consideration of a utility based mining approach was motivated by researchers due to the limitations of frequent or rare itemset mining which permits a user to suitably communicate his or her views regarding the usefulness of itemsets as utility values and then find itemsets with high utility values higher than a threshold. Identifying the lively customers of each such type of itemset mined and rank them based on their total business value can be done by these set of algorithms. This would be enormously supportive in developing Customer Relationship Management (CRM) processes like campaign management and customer segmentation. In all types of utility factors like profit, significance, subjective interestingness, aesthetic value etc the utility based data mining is a newly absorbed research area. This can add economic and business utility to existing data mining processes and techniques. A research area inside utility based datamining identified high utility itemset mining is intended to discover itemsets that introduce high utility.

Literature review: Identifying the association rules in large databases play a key role in data mining. Prasad and Ramakrishna (2011) have considered the prior researches and present working status in order to restore the gaps between them with present known information. There were two problems regarding this context: identifying all frequent item sets and to generate constraints from them. Here, first problem as it takes more processing time was computationally costly. Consequently, many algorithms were proposed to solve this problem. Their current study considers such algorithms and the related issues.

Saravanabhavan and Parvathi (2011) have presented an efficient tree structure for mining high utility itemsets. At first, they have developed a utility frequent-pattern tree structure an extended tree structure for storing crucial information about utility itemsets. Then, the pattern growth methodology was utilized for mining the complete set of utility patterns. Improved high utility itemsets mining efficiency was achieved using two major concepts: compressing a large database into a smaller data structure as well as the utility FP-tree avoids repeated database scans and the pattern growth method utilized in the proposed FP-tree-based utility mining avoids the costly generation of a large number of candidate sets and thereby reduces the search space dramatically. Experimental analysis was carried out on tree structure mining concept using different real life datasets. The performance evaluation results have demonstrated the efficiency of the proposed approach in mining high utility itemsets.

Association rules are the most important tool to discover the relationships among the attributes in a database. Prakash *et al.* (2011) have discussed that the existing association rule mining algorithms were applied on binary attributes or discrete attributes, in case of discrete attributes there was a loss of information and these algorithms take too much computer time to compute all the frequent itemsets. By using Genetic Algorithm (GA), it is possible to improve the generation of frequent itemset for numeric attributes. The major advantage of using GA in the discovery of frequent itemsets is that they perform global search and its time complexity was less compared to other algorithms as the genetic algorithm was based on the greedy approach. The main aim of their study is to find all the frequent itemsets from given data sets using genetic algorithm.

The main goals of Association Rule Mining (ARM) are to find all frequent itemsets and to build rules based of frequent itemsets. But, a frequent itemset only reproduces the statistical correlation between items and it does not reflect the semantic importance of the items. To overcome this limitation, Kannimuthu *et al.* (2011) have utilized a utility based itemset mining approach. Utility-based data

mining is a broad topic that covers all aspects of economic utility in data mining. It takes in predictive and descriptive methods for data mining. High utility itemset mining is a research area of utility based descriptive data mining, aimed at finding itemsets that contribute most to the total utility. The well known faster and simpler algorithm for mining high utility itemsets from large transaction databases is Fast Utility Mining (FUM). In this proposed system, they made a significant improvement in FUM algorithm to make the system faster than FUM. The algorithm was evaluated by applying it to IBM synthetic database. Experimental results have shown that the proposed algorithm was effective on the databases tested.

Sandhu *et al.* (2011) have proposed an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach has utilized traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of association rules mined were subjected to weightage (W-gain) and utility (U-gain) constraints and for every association rule mined, a combined utility weighted score (UW-Score) was computed. Ultimately, they have determined a subset of valuable association rules based on the UW-Score computed. The experimental results have demonstrated the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

Kuthadi (2013) has proposed an enhanced Association Rule Mining Algorithm to mine the frequent patterns. The algorithm utilized weightage validation in the conventional association rule mining algorithms to validate the utility and its consistency in the mined association rules. The utility is validated by the integrated calculation of the cost/price efficiency of the itemsets and its frequency. The consistency validation is performed at every defined number of windows using the probability distribution function assuming that the weights are normally distributed. Hence, validated and the obtained rules are frequent and utility efficient and their interestingness are distributed throughout the entire time period. The algorithm was implemented and the resultant rules were compared against the rules that can be obtained from conventional mining algorithms.

MATERIALS AND METHODS

The analysis of existing research works asserts that lack of dealing and drawbacks are present in existing data mining methods. In this research, I intend to develop a new efficient mining algorithm to mine the closed frequent patterns from the data base. During the closed frequent itemsets mining, a hash table is maintained to check whether the given itemset is closed or not. The computation of closed frequent itemsets from the data stream will minimize the memory usage and processing time. Since, set of frequent closed itemsets has smaller size rather than complete set of frequent patterns while it contains the same information. That is, the complete set of frequent itemsets can be induced by closed frequent itemsets. Therefore, closed itemset mining over data streams is more desirable than finding the complete set of frequent itemsets. The new closed itemsets mining algorithm will use probability distribution function in the frequency and utility methods. In frequency methods, the patterns or items that have high frequency rate are mined from the data base whereas in utility methods items with high profit rate are extracted. But, the proposed method will determine distribution of both highly frequent and profitable items. After determining the probability distribution of the items, items will be selected based on their distribution value. Items that have high distribution value will be selected because such items are likely to have same frequency and priority level in the future also. Thus, the proposed closed itemset mining algorithm will extract accurate patterns and overcomes the above mentioned problem. The proposed closed frequent itemset mining algorithm basic structural diagram is given in Fig. 1.

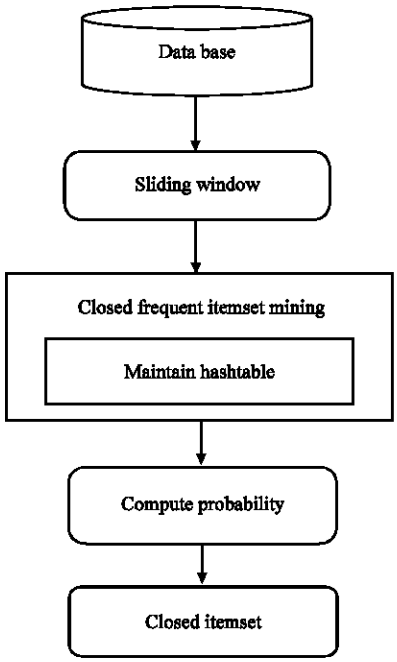

Fig. 1: Proposed structural diagram

RESULTS AND DISCUSSION

Motivation of the research: The recent research works are concisely reviewed in the previous unit. From the review, it can be seen that the previous research works have performed the data mining based on frequency and utility of the item sets. Many methods have been proposed for mining items or patterns from data base. These methods use frequency for extracting patterns from the data base. But, frequency based extraction is not always successful. In addition, frequency methods have some drawbacks. To overcome these drawbacks, the utility (priority) based method was introduced. Utility based methods extract patterns or items based on the weight or priority of the items. The individual performance of these methods over the history of data base mining has drawbacks. Accordingly, many works are developed using both frequency and utility methods and such works perform satisfactorily in mining items from the data base. But, these works do not provide assurance that the extracted patterns will continue to provide the same level of profit and frequency in the future. No literature work is available to solve this drawback. To overcome this problem, a mining algorithm which is given by Kuthadi (2013) was proposed for extracting patterns from data base using both frequency and utility methods.

But, this method has the drawback of memory usage and processing time. Because, in data streams data elements are arrive at a rapid rate. The incoming data is unbounded and probably infinite. Due to high speed and large amount ofincoming data, frequent itemset mining algorithm must require a limited memory and processing time. To reduce this drawback, a new algorithm is proposed in this study.

CONCLUSION

In this study, we present a deeper insight into the different data mining techniques that used to mine the significant information from the databases based on the patterns (or) item sets frequency or utility. Here all the existing methods are working efficiently but these techniques not considered the item sets frequency and utility in the future, i.e., these works do not provide assurance that the extracted patterns will continue to provide the same level of utility and frequency in the future. This study proposes a new approach for itemset mining on utility and frequency using distribution model. As a result, this study will be supportive for the researchers to improve the concentration on this data mining techniques by considering the item sets or patterns utility and frequency in the future also.

REFERENCES

Deogun, J., V. Raghavan, A. Sarkar and H. Sever, 1997. Data Mining: Research Trends, Challenges and Applications. In: Rough Sets and Data Mining: Analysis of Imprecise Data, Lin, T.Y. and N. Cercone (Eds.). Kluwer Academic Publishers, Boston, MA., USA., pp: 9-45.

Erwin, A., R.P. Gopalan and N.R. Achuthan, 2007. A bottom-up projection based algorithm for mining high utility itemsets. Proceedings of the 2nd International Workshop on Integrating Artificial Intelligence and Data Mining, Volume 84, December, 2007, Gold Coast, Australia, pp: 3-11.

Erwin, A., R.P. Gopalan and N.R. Achuthan, 2008. Efficient mining of high utility itemsets from large datasets. Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, May 20-23, 2008, Osaka, Japan, pp: 554-561.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM, 39: 27-34.

Jackson, J., 2002. Data mining: A conceptual overview. Commun. Assoc. Inform. Syst., 8: 267-296.

Kannimuthu, S., K. Premalatha and S. Shankar, 2011. iFUM-improved fast utility mining. Int. J. Comput. Applic., 27: 32-36.

Khalilzadeh, N. and P.J.M. Fard, 2008. Application of data mining in marketing and managing customer relationship. Proceedings of the 3rd International Marketing Management Conference, (MMC'08), Tehran, Iran, pp: 1-13.

Kuthadi, V.M., 2013. A new data stream mining algorithm for interestingness-rich association rules. J. Comput. Inform. Syst., 53: 14-27.

Malik, H.H. and J.R. Kender, 2007. Optimizing frequency queries for data mining applications. Proceedings of the 7th IEEE International Conference on Data Mining, October 28-31, 2007, Omaha, NE., USA., pp: 595-600.

Podpecan, V., N. Lavrac and I. Kononenko, 2007. A fast algorithm for mining utility-frequent itemsets. Proceedings of the 18th European Conference on Machine Learning and 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 21, 2007, Warsaw, Poland, pp: 9-20.

Prakash, V.R., Govardhan and S.S.V.N. Sarma, 2011. Mining frequent itemsets from large data sets using genetic algorithms. IJCA Special Issue Artif. Intell. Tech.-Novel Approaches Pract. Applic., 4: 38-43.

Prasad, K.S.N. and S. Ramakrishna, 2011. Frequent pattern mining and current state of the art. Int. J. Comput. Applic., 26: 33-39.

Sandhu, P.S., D.S. Dhaliwal and S.N. Panda, 2011. Mining utility-oriented association rules: An efficient approach based on profit and quantity. Int. J. Phys. Sci., 6: 301-307.

Saravanabhavan, C. and R.M.S. Parvathi, 2011. Utility FP-tree: An efficient approach to mine weighted utility itemsets. Eur. J. Scientif. Res., 50: 466-480.

Velickov, S. and D. Solomatine, 2000. Predictive data mining: Practical examples. Proceedings of the 2nd Joint Workshop on Artificial Intelligence Methods in Civil Engineering Applications, March 26-28, 2000, Cottbus, Germany, pp: 1-16.

Yao, H., H.J. Hamilton and L. Geng, 2006. A unified framework for utility-based measures for mining itemsets. Proceedings of the Second Workshop on Utility-Based Data Mining, August 20, 2006, Philadelphia, PA., USA., pp: 28-37.