Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (66-69), Month: May - August 2015, Available at: www.noveltyjournals.com

MINING INFREQUENT ITEMSETS USING APRIORI ALGORITHM

[1]Ms.L.Jayachitra, [2]Mrs.S.Lakshmi Sridevi

[1]M.Phil student, Department of MCA, Hindustan University, Padur, Chennai
[2]Assistant Professor, Department of MCA, Hindustan University, Padur, Chennai

Abstract: **The aim of Association Rule Mining is to find the correlation between data Items based on frequency of occurrence. Infrequent Item set mining is a variation of frequent item set mining where it finds the uninteresting patterns i.e., it finds the data items which occurs very rarely. Considering weight for each distinct item in a transaction independent manner adds effectiveness for finding frequent item set mining. Several articles related to frequent and weighted infrequent item set mining were proposed. This paper focus on reviewing various Existing Algorithms related to frequent and infrequent item set mining which creates a path for future researches in the field of Association Rule Mining.**

Keywords: **Infrequent Item set mining, Association Rule Mining, weight, Correlation.**

1. INTRODUTION

1.1 Introduction:

Knowledge Discovery and Data Mining (KDD) is playing an important role in extracting knowledge in this era of data overflow. KDD consists of many methods and techniques that can be applied to different data to extract knowledge. Some of the methods include association, classification, and clustering [14].

Association rule mining is the discovery of association relationships among a set of items in a dataset.

Association rule mining has become an important data mining technique due to the descriptive and easily understandable nature of the rules. Although association rule mining was introduced to extract associations from market basket data [14], it has proved useful in many other domains (e.g. microarray data analysis, recommender systems and network intrusion detection). In the domain of market basket analysis, data consists of transactions where each is a set of items purchased by a customer. A common way of measuring the usefulness of association rules is to use the support-confidence framework introduced by [14].

Support of a rule is the percentage of transactions that carry all the items in the rule, and the confidence is the percentage of the transactions that carry all the items in the rule among those transactions that carry the items in the antecedent of the rule.

The problem of association rule mining can be stated as: Given a dataset of transactions, a threshold support (minsupport), and a threshold confidence (minconfidence); Generate all association rules from the set of transactions that have support greater than or equal to minsupport and confidence greater than or equal to minconfidence.

1.2 Association Rule:

Initially it was largely motivated to understand the market basket data, the results of which allowed companies to understand purchasing behavior and, as a result, better target market audiences. ARM is user centric as the objective is the elicitation of interesting rules from which new knowledge can be derived. ARM is to facilitate the discovery, heuristically filter, and enable the presentation of these inferences or rules for subsequent interpretation by the user to determine their usefulness. ARM has been divided into two phase of process as follows:

Phase 1: Identify the sets of frequent items or item sets or pattern within the set of transaction using user-specified support threshold.

Phase 2: Generate inferences or rules from these above patterns using user-specified confidence threshold.

The above two phases are generated strong association rules from dataset. The first phase is called frequent item set construction or mining. That is extremely computational expensive than phase 2. The second phase is called association rule generation. That is, straight forward process. This phase computational complexity is negotiable to compare with first phase. There are two major problems in second phase. The first problem is rule quantity means that algorithms can produce large number of rules. The second problem is rule quality means that, all the rules are not interesting. The support and confidence measures play a vital role to filter unwanted item sets and rules from the mining process.

1.3 Types of Association Mining:

1.3.1 Positive Association Rule Mining:

The classical association rules consider only items enumerated in transactions of the dataset. The positive relationship can be found between the set of items. The rules are generated from the positive related items. These rules are referred to as positive association rules. Most of the algorithms were developed for generating positive associations between items. These are useful to decision making [16].

The positive rules are classified as follows:

1. Boolean association rule

a. Quantitative [1]

b. Constrained rules [3]

c. Sequential rules [4]

2. Qualitative association rule [1]

3. Spatial association rule

4. Temporal association rule

1.3.2 Negative Association Rule Mining:

Negative association rules also consider the same items, but in addition the item also considers which were absent from transactions. The negative rules are generated from infrequent item sets. These rules play some important role in decision-making [16]. These are useful in market basket analysis to identify products that conflict with each other or products that complement each other. This is a difficult task, due to the fact that there are essential differences between positive and negative rule mining.

Brin et al [5] mentioned for the first time in the literature the notion of negative relationships.

The authors have used statistical chi-square test to verify the independence between two variables. The authors have also used correlation measure to determine the nature (positive or negative) of the relationship. The strong negative rules are mined by Savasere et al [6]. They combined positive frequent item sets with domain knowledge in the form of taxonomy.

1.3.3 Constraint based Association Rule Mining:

The constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the rules. By doing this lots of cost of mining those rules that turned out to be not interesting can be saved. Usually constraints are provided by users. The constraints are classified as follows:

1. Knowledge based constraints [7]

2. Data constraints [8]

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (66-69), Month: May - August 2015, Available at: www.noveltyjournals.com

2. SYSTEM

2.1 Existing System:

In the existing system the Apriori algorithm was used to find the frequent item sets. For finding the frequent item sets in the medical database makes more time to spend the treatment with the healthier people.

So doctor's time was wasted in the existing system. So we are moving to find the infrequent data in the proposed system.

2.2 Proposed System:

In the proposed algorithm is finding infrequent item sets using the Apriori property is applicable that is if we find an item set as infrequent then all its supersets are considered as infrequent but here these super sets are not to be pruned as in Apriori and they are considered in to the solution as infrequent k- item sets.

Since the proposed system uses medical application to find the infrequent diseases in order to give the high response to abnormal diseases.

In the study of finding a better treatment approach for a special disease, researchers would like spend more time on studying an abnormal case rather than reading the millions of records of healthy people. In this scenario, more effort has been put into the development of infrequent item set mining.

3. CONCLUSION

Since the late 1990s, more and more researchers have realized the importance of infrequent patterns with the increasing demands from applications of anomaly detection, especially in medicine, genetics, molecular biology and network security.

4. FUTURE WORK

The proposed algorithm works efficiently to find the infrequent item sets. It finds all the infrequent item sets within one data base scan. But the proposed method not considers any pruning strategy. It is better to implement any pruning strategy to improve the complexity of the proposed method.

REFERENCES

[1] Srikant, R and Agrawal, R. "Mining quantitative association rules in large relational tables." Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, pp.1-12.

[2] Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993ACM- SIGMODInternational conference on management of data (SIGMOD'93), Washington, DC, pp 207– 216.

[3] Padmanabhan, B., Tuzhilin, A. 1998. "A belief-driven method for discovering unexpected patterns." Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York city, USA, August 27-31, pp.94-100.

[4] Agarwal, R., Srikant,R. 1995. "Mining sequential patterns." Proceedings of Eleventh International Conference on Data Engineering, Taipei, Taiwan, March 6-10, pp.3-14.

[5] Brin, S., Motwani, R., Silverstein, C. 1997. "Beyond Market Baskets: Generalizing Association Rules to Correlations." Proceedings of the ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, May 13-15, pp.265-276.

[6] Savasere, A., Omiecinski, E., Navathe, S. 1998. "Mining for strong negative associations in a large database of customer transactions." Proceedings of the International Conference on Data Engineering, Oralando, Florida, USA, February 23-27, pp.494-502.

[7] Ng, R. T., Lakshmanan, L. V. S., Han, J., Pang, A. 1998. "Exploratory mining and pruning optimizations of constrained association rules." Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA, June 2-4, pp.13-24.

[8] Bayardo, J.R., Agarwal, R., Gunopulos, D. 1999. "Constraint-based rule mining in large, dense databases." Data Mining and Knowledge Discovery Journal, Vol.4, Issue:2-3, pp.217-240.

[9] Salleb, A., Turmeaux, T., Vrain, C., and Nortet, C. 2004. "Mining quantitative association rules in a atherosclerosis dataset." Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, pp.98-103.

[10] Ordonez, C., Ezquerra, N., Santana, C.A. 2006. "Constraining and summarizing association rules in medical data." International Journal of Knowledge Information System, Vol.9, Issue.3, pp.259-283.

[11] Ordonez, C., Santana, C.A., Braal, L. 2000. "Discovering interesting association rules in medical data." Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Dallas, Texas, USA, May 14, pp.78-85.

[12] Gasmi, G., Hamrouni, T., Abdelhak, S., Ben Yahia, S., Mephu Nguifo, E. 2005. Extracting generic basis of association rules from SAGE data, Proceedings of the Eighth International ECML/PKDD Workshop Discovery Challenge, Porto, Portugal, October 7, pp.1-6.

[13] Chiu, H.W., Hung, F.H. 2008. "Association Rule Mining from Yeast Protein Interaction to Assist Protein-Protein Interaction Prediction" Biomedical Soft Computing and Human Sciences, Vol.13, No.1, pp.3-6.

[14] R. Agrawal, T. Imieinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, Proceedings of the ACM SIGMOD International Conference on the Management of Data, pages 207– 216, Washington DC, 1993. ACM Press.

[15] Gupta, N., Mangal, N., Tiwari, K., Mitra, P. 2006. "Mining Quantitative Association Rules in Protein Sequences." Data Mining, Lecture Notes on Artificial intelligence 3755, Springer-Verlag, Berlin, pp.273-281.

[16] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems, 22(3):381–405, 2004.

[17] Laxminarayan, P., Ruiz, C., Alvarez, S.A., Moonis, M. 2005. "Mining Associations over Human Sleep Time Series." Proceedings of the eighteenth IEEE Symposium on Computer-Based Medical Systems, Dublin, Ireland, June 23-24, pp.323-328.

[18] Chen, Q., Chen, Y. 2006. "Mined frequent patterns for AMP-activated protein kinase regulation on skeletal muscle." BMC Bioinformatics, Vol.7, No.394, pp.1-14.

[19] Ronaldo Cristiano Prati, Maria Carolina Monard and André C.P.L.F. de Carvalho. 2004. "Looking for exceptions on knowledge rules induced from HIV cleavage data set." International Journal Genetics and Molecular Biology, Vol.27, Issue.4, pp.637-643.

[20] Sengul Dogan., Ibrahim Turkoglu. 2008. "Diagnosing hyper lipidemia using association rules." Mathematical and Computational Applications, Vol. 13, No. 3, pp.193-202.

[21] Shantakumar B.Patil., Kumaraswamy, Y.S. 2009. "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction." International Journal of Computer Science and Network Security, Vol.9 No.2, pp.228-235.

[22] Agrawal, Rakesh; and Srikant, Ramakrishnan; Fast algorithms for mining association rules in large databases.