International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

A Novel Approach to generate Bit-Vectors for mining Positive and Negative Association Rules

G. Mutyalamma [1] , K. V Ramani [2] , K.Amarendra [3]

[1] M.Tech Student, Department of Computer Science and Engineering, Dadi Institute of Engineering & Technology, Anakapalle,
[2] Sr.Asst.Prof Department of Computer Science and Engineering, Dadi Institute of Engineering & Technology, Anakapalle,
[3] Professor, Department of Computer Science and Engineering, Dadi Institute of Engineering & Technology, Anakapalle, A.P, INDIA.

Abstract The main motto of any data mining process is to extract reports by the aid of semi-automatic or automatic process of analyzing large quantitative patterns from heterogeneous data sets. When large sets are co-related, well researched methodologies are required for generating strong association rules. These rules are posed depending on support or confidence measures of significance and interestingness with respect to minimum thresholds. Multi-objective heuristic algorithms are to be proposed for deriving patterns which falls under positive and negative associations. Many of the basic algorithms existing are purely derived for positive association rules by considering only one evaluation criteria, but recent data sets demand for populating negative associations for better understandability. This paper proposes a novel method for generating bit vectors for positive and negative associations. Comprehensibility, interestingness and performance are maximized by reducing the time complexity for frequent and infrequent item set generation of positive and negative association rules with more flexibility.

***Index Terms*— bit vectors, infrequent item-sets, negative association, support.**

Introduction

Data mining has been vital in terms of discovering patterns from large set of databases. Data mining has been coined as an intermediate step of knowledge discovery of data. It has been a blended of many a process for extracting reports on par of interestingness of the target user. Data mining can be defined as an interdisciplinary approach for mapping data sets and to the process of visualizing it. It is a process of transforming raw data into its understandable format meant for future usage.

Data mining aims at analyzing data into set of data groups; termed as clusters; or as a set of unusual dependencies ; termed to be outliers in analysis ; or as a set of dependencies. The set of dependencies that incur between any sets of data is termed as a process of association. Association has been a prominent endeavor for analyzing dependency of one data object on the other, which is generally associated by means of support and confidence. This association study has become prominent in many disciplines like market basket analysis, fraud detection in password management and in many decision support systems, intrusion detection, and telecommunication.

Association rule mining is associated in deriving frequent features termed as frequent item sets. Positive association rules are generated in visualizing and predicting the outcomes by analyzing the support and confidence factors. Many algorithms have been coined for analyzing and generating rules depending upon the level of association that is hailing between data objects. The rules generated are maintained for future prediction analysis and many years the data sets with minimal support are simply ignored or pruned as they form negative associations.

Decision support system is build only on the basis of positive association rules generated. Recent researches have proved that the negative associations which deal with infrequency in item set

generations are also important for analyzing the robustness of the system and to build a reliable system. Mining of positive and negative associations has attained demand in studying frequent and infrequent item sets. Much effort is to be posed for analyzing negative associations. These associations are used for extracting frequent items from infrequent item sets and vice-versa by minimizing the threshold levels of associations. Highly correlated data objects are analyzed with ease. Many frameworks have been designed for maintaining such infrequency in item set generation. Additional interesting measures are generally added to reduce the negative associations or same measures are considered for framework for deriving robust association rules.

Most of the Association rules generally rely on single evaluation criteria, termed as mono-objective algorithm with certain limitations of optimizing multi-interesting measures for easy understanding and good coverage of the data set objectives. Recent researches are been proposed for extracting association rules as a multi-objective, by considering several objectives in the process of extracting associations depending on interestingness.

Bit vector generation has been proposed for generating quantitative association rules of positive and negative association with much reduced time complexity and with more flexibility. Bit vectors are derived depending on the occurrence of data items. Care is taken for analyzing positive and negative Boolean associations with thresholds defined to support and confidence measures. Reformation of associations rules are performed as per the scalability of the data sets and threshold level measures depending upon interestingness.

Literature review

Research has been made for extracting information from huge set of data repositories. Data Mining has been a vital process of extracting useful patterns. Search engines are been developed for extracting useful patterns. Mining is made an inter-disciplinary process of knowledge discovery. User-centric reports are generated for unknown data patterns as per business requirements [1].

Mobility of data has made the process of extracting information, a complex task as data comprises of different sources, structures and ownerships. Scalability is been abundantly increased and process of report generation has become hectic with the increase in cross-functional aspects. Association discovery is termed to be one of the primitive and common methodologies in extracting knowledge [2]. Generating association rules has become a common process for analyzing the dependency between item sets primitively on binary or discrete values. Antecedent and consequent relations have been derived for item sets generated from huge datasets, depending upon the interestingness in the patterns. Recent studies have proved that data tends to be quantitative and hence quantitative association rule mapping (QAR) has attained attention and interest [3] [4].

Apriori algorithm with basic steps of frequent item set generation and association rule generation has become the most well-known process of association rule mining. The increase in the demand of analyzing frequency of patterns has risen to a new methodology of finding frequent item set generations with the aid of FP-Tree generation. This process, termed as FP-growth makes analyzing of huge data with limited number of scans [5]. Most of the algorithms proposed are intended only for positive associations and the introduction to the fuzzy approach has made the process of generalizing associations on the basis of absent item sets for generating rule dependency measures [6].

Apriori Algorithm generally suffers with “Rare item set generation” problem where associations are missed for certain sensitive and rare items, when there is a very low minimum support value generated. This even suffers with generating missing rules or explosion rules among data sets. A new variant has been defined for analyzing data item sets with multiple values for the minimal support value for generating frequent or infrequent item set rules. This approach is termed as MSApriori, Multiple Minimum Support Apriori, which still suffers with “rear item problem”, when there is large variance in support [7].

DI-Apriori algorithm has been retrieved to study reverse associations. These tend to extract dissociation rules for negatively associated data sets, where the number of patterns generated is low [8]. Simultaneous positive and negative association rule mining is also proposed as other variant of Apriori which purely depends on Pearson correlation, basing on the level of interest measure that one uses at prune step [9]. Depending on the correlations one positive rule for strong correlation and two negative rules for infrequent item sets are defined. New frameworks have been developed deriving confined negative associations where antecedent and consequent is a combination of conjunctions of either negated or non-negated attributes [10].

Multi-objective evolutionary algorithms (MOEA) are derived for mining QAR with varying trade-offs in evaluating criteria. Recent MEOA, termed as MEOA/D, which are MEOA depending on decomposition are used to explicitly decompose multi-objective optimization problem into N scalar optimization sub-problems and to optimize them simultaneously [11] [12].

Other variant has been derived to generate positive and negative associations, PNQAR, dealing with strong and weak associations that exist between different data objects. This helps in attaining a good trade-off between rules, support and coverage with respect to user perspective [13].

In order to mine positive and negative QARs with much less computational cost, for maximizing basic objectives like comprehensibility, interestingness and performance, a new MOEA algorithm, MOPNAR is derived in order to increase the population of coverage of datasets with the help of external population (EP). MOPNAR, when compared to PNQAR is used for generating association rules with less computational effort with good scalability and low computational cost [14].

Methodology

In general, association rules are derived in terms of discrete or binary values depending on the interestingness measure. Patterns are derived either foe a mono-objective or multi-objectives defined by the user. Associations generally claim to have strong positive correlations between data objects in the data sets. Certain conditions prompt to consider negative associations for attaining sensitive and sensible rules of correlation between the data objects. Fuzzy type correlations can be derived between objects for attaining information about the existence or absence of relations. Bit vectors are generated for storing data presence in the data sets depending on which associations are being derived. Algorithms are being designed for both positive and negative associations from both of the frequent and infrequent item sets that are generated. This is accomplished by the aid of algorithms which are meant for analyzing frequent and infrequent item sets.

Algorithms are coined for

1. Analyse and derive frequent and infrequent item sets.
2. Define strong positive and strong negative associations on the item sets derived depending upon the minimal support value that is considered.

Though the process of defining positive and negative associations seems to be the core aspect of mining, but process of initial selection of frequent and infrequent item sets on par of interestingness is more complex task that is to be performed.

Boolean associations are being derived from the database of transactions, in terms of user defined minimal support and confidence values. The task of identification of frequent and infrequent item sets are accomplished with initial step of deriving Bit-Vectors for deriving Quantitative association rules for both positive and negative associations from the transaction database.

The process of generation of Bit-Vectors is a blended of two prominent phases- generation of BV and the generation of frequent and infrequent item sets from the transaction databases.

Phase 1: Generation of Bit-Vectors

1. Examine the entire item sets from the transactional database that is considered.

2. Divide the items depending on existence and store item based vectors, generated into BV, for all items in the transactional database.

3. Examine whether a particular item exists in the database, if it exists, then place a value of one in the BV factor.

4. In case if the item does not exist, then mark items to be infrequent and denote tem with a value of zero in the BV.

5. Repeat steps 3 and 4 for all data items that exist in the transaction database.

Phase 2: Frequent and Infrequent item sets generation.

1. Assign index for frequent item sets of factor 1, to FreqIndexFact.

2. Candidate item sets are to be generated for k–items from the frequent item sets derived. For each I in candidate of K-items, define CS_K.

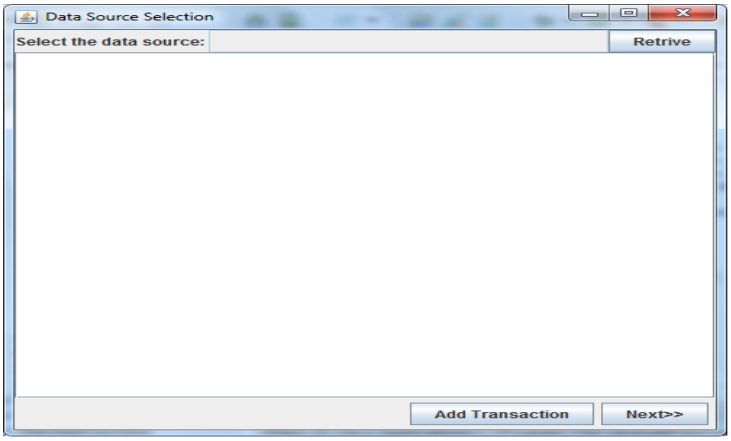
3. If support (I) ≥ minimal_support and Correlation (I)>1, assign I to frequent K-item set list (FIL_K), else assign the value to IFL_K which denotes Infrequent item-sets for K-item list.

4. Derive support value for each I in the transactional database by XOR operation between the bit vectors values generated. In case if I_1 and I_2 are the bit vectors generated, and ¬I is the association derived, then the support value is derived as support(I) = I_1 ^ I_2 .

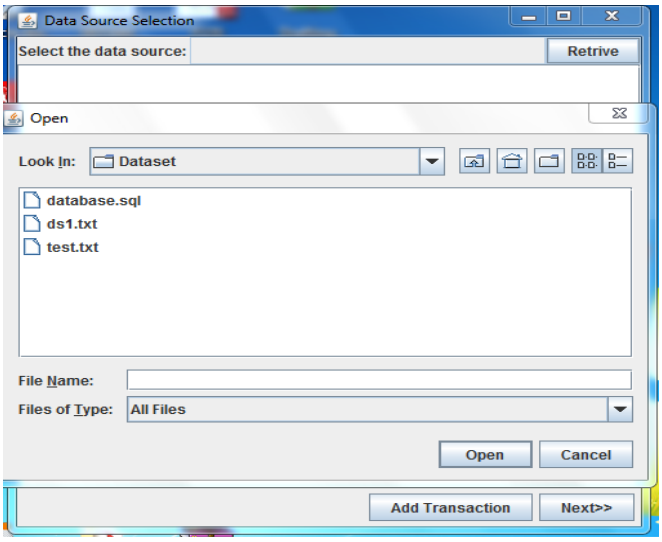
5. Now assign the values derived to Frequent and Infrequent lists with respective indices foe k values. i.e.; assign FIL_K and $IFIL_K$ to FreqIndexFact and InFreqIndexFact for all K-item sets respectively.

Experimentation and Results

To derive the positive and negative associations, using bit vectors; initial step has been taken in defining an input data set for which the bit vectors are to be defined. Bit vectors are then used for deriving positive and negative associations depending on the interestingness of the patterns. A sample data set has been designed with certain character value on which bit vectors are generated for mining infrequent sets. Code has been developed to prompt the data worker to select a particular data set.



The above figure clearly depicts the cover screen which prompts the user to select the data source for processing.



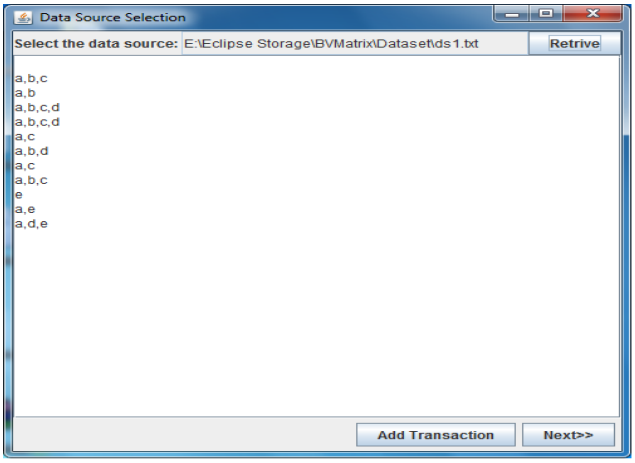
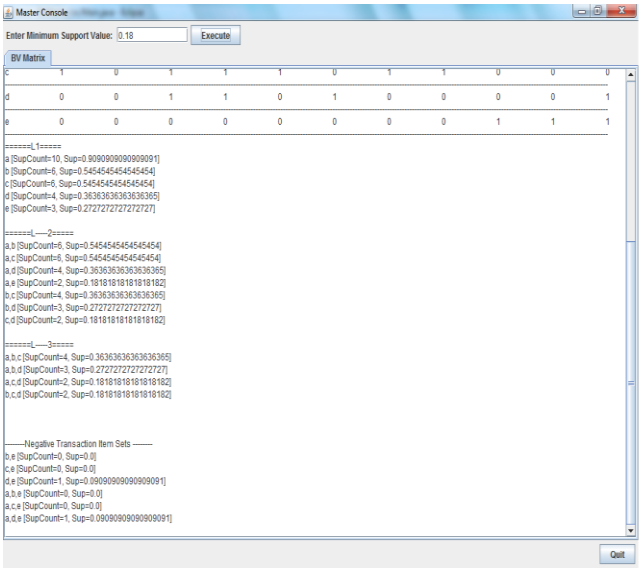


Figure clearly shows the input sequences that have been taken for analyzing the infrequent item sets. Data set taken is a synthetic data with some stereotypes for data sets that forms negative associations.



When the particular data sets have been selected, first bit vectors have been derived depending upon the support values for one item to occur in the presence of the other. Minimum support count has been optimally chosen and the support value is derived for n-set item-sets. Figure shows the support values for different item set levels. Negative item sets are derived depending on the bit vectors of the transactional data sets. In this case, the support values are generated when the minimum support value is taken to be 0.18. The above figure clearly depicts that the combinational character patterns like {b, e}, {c, e}, {a, b, e} and {a, c, e} forms infrequent sets with support value and support count to be zero.

Conclusion

Mining of negative associations has got a greater demand, which is termed to be highly relevant and at the same time, to be highly hectic in computing. Even associations are generally used for analysing user interestingness, which are termed to be frequent. Infrequent item sets are also vital in deriving nominal reports. Many of the primitive algorithms have derived infrequent sets for certain extent, but this proposed system derives the negative associations by deriving bit vectors. This proposed system has made the concept of utilizing these bit vectors, initially derived. Minimum support is defined and support count for each of the pairs of associations between bit vectors is derived.

VI. References

[1] Anuradha Srinivas , Practical approach towards data mining and its analysis (2013).

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2006.

[3] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Proc. ACM SIGMOD*, 1996, pp. 1–12.

[4] K. Sun and B. Fengshan, "Mining weighted association rules without pre assigned weights," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 4, pp. 489–495, Apr. 2008.

[5] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACMSIGMOD Record*, vol. 29, no. 2, pp.1–12, 2000.

[6] M. Delgado, M. D. Ruiz, D. S´anchez, and J. M. Serrano, "A formal model for mining fuzzy rules using the *RL* representation theory," *Information Sciences*, vol. 181, no. 23, pp. 5194–5213, 2011.

[7] R. Uday Kiran and P. Krishna Reddy" An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules "978-1-4244-2765-9/09 © IEEE 2009.

[8] M. Morzy, "Efficient mining of dissociation rules," in *Data Warehousing and Knowledge Discovery*, pp. 228–237, Springer, 2006.

[9] P. H´ajek, I. Havel and M. Chytil,"The GUHA Method of Automatic Hypotheses Determination", *Computing*, Vol. 1, 1966, pp 293–308.

[10] M.L. Antonie and O.R. Za¨ıane,"*Mining Positive and* Negative Association Rules: an Approach for Confined *Rules"*, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2004, pp 27–38.

[11] D. Martin, A. Rosete, J. Alcala-Fdez, and F. Herrera, "A multi objective evolutionary algorithm for mining quantitative association rules," in *Proc. 11th Int. Conf. Intell. Syst. Design Applicat.*, Nov. 2011, pp. 1397–1402.

[12] H. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi-objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Syst. Applicat.*, vol. 38, no. 1, pp. 288–298, 2011.

[13] A. Eiben and J. Smith, *Introduction to Evolutionary Computing*. Berlin, Germany: Springer-Verlag, 2003.

[14] Diana Mart´ın, Alejandro Rosete, Jes´us Alcal´a-Fdez, and Francisco Herrera, New Multi-objective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules, IEEE Transactions on Evolutionary Computation, Vol.18, No.1, February-2014.