Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.5 – 8

SURVEY ON EFFICIENCY OF ASSOCIATION RULE MINING TECHNIQUES

[1]MS.J.OMANA, [2]MS.S.MONIKA, [3]MS.B.DEEPIKA

[1]Assistant Professor- I, Information Technology, Prathyusha Engineering College, Chennai, India
[2,3]Student, Department of Information Technology, Prathyusha Engineering College, Chennai, India

ABSTRACT: Association rule mining is the technique that can discover set of frequent items in a transaction. It also helps in generating ruleset. These rulesets are the fundamental part of machine learning process. Here we discuss about APRIORI, ELCAT and OPUS. This paper aims at discussing the process carried out in the association rule mining techniques also the merits, demerits and applications of the algorithm.

KEYWORD: Association Rule Mining, Apriori, Eclat, Opus

I. INTRODUCTION

Data Mining is the process of discovering knowledge. It is the process of extracting information from available raw data. The data are stored in databases. There are various kinds of data that can be used in data mining which includes transactional data, statistical data etc. Data mining includes various techniques for each purpose. Techniques include Association rule mining, classification and prediction, regression etc.

Association rule mining techniques are widely used in discovering hidden correlations and relationship between set of items in a transaction. It includes every transaction in the database during the discovery process. It also reveals the set of strategies that can be followed or neglected in the field for respective development.

II. RELATED WORKS

[4] Comparative Survey on Association Rule Mining Algorithms. Manisha Girotra, Saloni Minocha, Kanika Nagpal and Neha Sharma. International Journal of Computer Applications, December 2013. In this paper, the algorithms that use association rules are divided into two stages, the first is to find the frequent sets and the second is to use these frequent sets to generate the association rules. It also provides a comparative study of different association rule mining techniques stating which algorithm is best suitable in which case. [7] Association Rule Mining :A Survey. Gurneet Kaur. International journal of computer science and information technologies. ARM technique along with the recent related work that has been done in this field. Techniques discussed is Apriroi. The paper also discusses the issues and challenges related to the field of association rule mining. [8] Comparative Analysis of Association Rule Mining Algorithms Based on Performance Survey. K.Vani. International journal of Computer Science and Information Technologies,2015. In this paper includes performance survey of Apriori, FP growth and eclat algorithms and compare those algorithms based on execution time using various datasets and support values. This paper also compared the merits and demerits of ARM algorithms.

III. ASSOCIATION RULE MINING

The set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analysed for buying patterns that reflect items that are frequently *associated* or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule.

Computer =>antivirus software [*support* = 2%; *confidence* = 60%]

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (5.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items.

ASSOCIATION RULES

- Implication: $X \rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \varnothing$;
- Support of AR (s) $X \rightarrow Y$:
  - Percentage of transactions that contain $X \cup Y$
  - Probability that a transaction contains $X \cup Y$.
- Confidence of AR (a) $X \rightarrow Y$:
  - Ratio of number of transactions that contain $X \cup Y$ to the number that contain X
  - Conditional probability that a transaction having X also contains *Y*.

A. APRIORI

Apriori is an association rule mining technique which when given the input of transactional database it mines all frequently occurring items in the transaction. Here when given the Electronic Medical Record as the input to Apriori it then generates a set of risk factors that occur frequently and indicates those to be factors for developing diabetes.

Pseudo-code

C_k: Candidate itemset of size k

L_k : frequent itemset of size k, L_1 = {frequent items};

for (k = 1; L_k !=Æ; k++) **do**

- C_{k+1} = candidates generated from L_k;
- **for each** transaction t in database **do**
  increment the count of all candidates in C_{k+1} that are contained in t;

endfor;

- L_{k+1} = candidates in C_{k+1} with min_support

endfor;

return $\cup_k L_k$;

Apriori works in the principle of support and confidence. The algorithm recursively generates candidate itemsets for each n-itemset. Apriori is capable of generating a very large set of risk factors and these can be used in prediction. Apriori is a recursive algorithm which extracts only one risk factor at a time and group them to produce the rule for prediction.

It generates a very large set of ruleset. The limitation is that it generates candidate itemset at each level of processing.

B. ELCAT

Elcat procedure is similar to that of the Apriori algorithm which functions in a recursive manner. It makes use of tree like structure known as the Tidset. Further it processes the given EMR to generate the frequently occurring risk factors of diabetes. The tidset starts with all one time occurring risk factor in the database.

The algorithm perform search in Depth first search manner to determine all the frequently occurring risk factor. The algorithm functions recursively and makes use of join operation() in combining all possible set of risk factors to the tidset and to generate (n+1) risk factors by considering the n^{th} itemset.

Eclat helps in lowering the memory being used during processing.

As it is similar to that of apriori algorithm, elcat also generates a very large set of ruleset but there is no candidate generation. The limitation is that it can discover only frequent item pattern rather ruleset is not generated.

C. OPUS

Opus is an efficient technique that functions recursively with respect to the parameters in the Left hand side and the right hand side. The algorithm considers the current left hand side, compares with available left hand side then updates the current left hand side. The parameters can be constrained for left hand side and the right hand side.

Unlike other association algorithms it monotonic doesn't require any parameter like support or confidence for Association rule mining. User can specify the maximum number of associations to be generated.

*current LHS: currently considered LHS of the rule
*available LHS: set of risk factors that can be added to the LHS of the rule

IV. CONCLUSION

In this paper, algorithmic aspects of association rule mining are dealt with. From a broad variety of efficient algorithms, the most important ones are compared. The algorithms are systemized and their performance is analysed based on runtime and theoretical considerations. Despite the identified fundamental differences concerning employed strategies, runtime shown by algorithms is almost similar. The comparison table shows that the Apriori algorithm outperforms other algorithms in cases of closed item sets whereas Eclat takes the lead in free item sets. Recursive Elimination was better than Apriori in all the cases but lacked in comparison to Eclat in all the cases. OPUS displayed better performance in all the cases leaving Eclat and Apriori behind by abolishi ng the concept of candidate generation and it does not require any support or confidence values to be mentioned during processing. The paper would give a basic idea to the company's data mining team about the algorithm which would yield better results.

REFERENCES

[1] A Survey: On Association Rule Mining. Jeetesh Kumar Jain, Nirupama Tiwari and Manoj Ramaiya. International Journal of Engineering and Research and Application. February 2013.
[2] An Extensive Survey on Association Rule Mining Algorithms. Mihir R Patel and Dipak Dabhi. International journal of Emerging technology and Advanced Engineering, January 2015.
[3] A Survey on Association Rule Mining. T.Karthikeyan and N.Ravikumar. International Journal of Advanced Research in Computer and Communication Engineering, January 2014.
[4] Comparative Survey on Association Rule Mining Algorithms. Manisha Girotra, Saloni Minocha, Kanika Nagpal and Neha Sharma. International Journal of Computer Applications, December 2013.
[5] Algorithms for Association Rule Mining- A General Survey and Comparison. Jochen Hipp,Ulrich Guntzer, Gholamreza. ACM SIGKDD, July 2000.
[6] Foto Afrati, Aristides Gionis, and Heikki Mannila. Approximating a collection of frequent sets. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2004.
[7] Assocaiation Rule Mining: A Survey. Gurneet Kaur. International journal of computer science and information technologies.
[8] Comparative Analysis of Association Rule Mining Algorithms Based on Performance Survey. K.Vani. International journal of Computer Science and Information Technologies, 2015.

