International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M32-072014

PERFORMANCE EVALUATION OF ASSOCIATION RULE MINING ON PARTITIONED DATA

Jagriti Ashri
M.Tech. Research Scholar
Ambala College of Engineering and Applied Research
Mithapur, Ambala Cantt., Haryana, India

Parneet Kaur
Assistant Professor
Ambala College of Engineering and Applied Research
Mithapur, Ambala Cantt., Haryana, India

ABSTRACT

Association rule mining is one of the major domain in data mining that is one of the excellent researched paradigms for the discovery of interesting and desired relations between variables from large databases. In many applications, it is required to identify strong and efficient rules in the databases using assorted measures of interestingness. This approach or paradigm is used widely in number of domains including banking, finance, market basket analysis, agriculture and many others. The kind of information fetched as association rules can be used as the base for decisions about marketing activities such as promotional pricing or items placements. Moreover, from market basket analysis, the association rules can also be employed in intrusion detection, Bio-Statistics, Continuous production and bioinformatics. In contrast to the sequence mining, association rule learning typically do not considers the sequence of products within a transaction or across multiple transactions. In this manuscript, the proposed research implemented the Apriori Algorithm as base algorithm for investigation on multidimensional database relations and finally the results are optimized using empirical algorithmic approach. Additionally, the formation and the prediction of the Association Rules of the Items Patterns shall be implemented by proposing the improvements in the using fuzzy logic. The simulation and implementation has been performed and the results extractions are interpreted for the efficiency analysis of the proposed research work. The proposed version of the Apriori algorithm alongwith the optimization module processing is proved to be efficient in terms of the execution time and obviously the complexity and phase reduction that will lead the proposed technique efficient as compared to the classical approach. This work mainly focuses on the performance of multidimensional database and its association with the rule mining.

Keywords

Multidimensional data mining, Apriori Algorithm, Association Rule Learning

1. INTRODUCTION

In today's growing world in assorted technologies, the role of data mining [1] is escalating day by day with the new aspect of business. Data mining has been proved as a very basic tool in knowledge discovery and decision making process. Data mining technologies are very frequently used in a variety of applications. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters. Frequent patterns are the itemsets that are frequently visited in database transactions at least for the user defined number of times which is known as support threshold. Presently a number of algorithms have been proposed in literature to enhance the performance of Apriori Algorithm, for the purpose of determining the frequent pattern.

2. KNOWLEDGE DISCOVERY

Data mining is the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. Knowledge discovery in databases (often called data mining) aims at the discovery of useful information from large collections of data. In addition the author puts special stress on fact that the task of knowledge discovery is inherently interactive and iterative, and it is a process containing several steps where data mining is one of them.

According to Hedberg, KDD is abbreviation of knowledge discovery and data mining, which may lead to confusion. The most sophisticated definition is one according to [2], where authors have determined that knowledge discovery in databases is interactive and iterative process with several steps and data mining is a part of this process. Process of KDD is defined as: The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Data mining is the process of applying these methods with the intention of uncovering hidden patterns [3] in large data sets. Data mining requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis. This may be supported by visualization tools, that display data, or through fundamental statistical analysis, such as correlation analysis. Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation. Versatile refers to the ability of the tool to apply a wide variety of models. Scalable tools imply that if the tools works on a small data set, it should also work on larger data sets. Automation is useful, but its application is relative. Some analytic functions are often automated, but human setup prior to implementing procedures is required. In fact, analyst judgment is critical to successful implementation of data mining. Proper selection of data to include in searches is critical. Data transformation also is often required. Too many variables produce too much output, while too few can overlook key relationships in the data. Fundamental understanding of statistical concepts is mandatory for successful data mining.

In today's growing world in assorted technologies, the role of data mining is escalating day by day with the new aspect of business. Data mining has been proved as a very basic tool in knowledge discovery and decision making process. Data mining technologies are very frequently used in a variety of applications. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases,

such as association rules, correlations, sequences, episodes, classifiers, clusters. Frequent patterns are the itemsets that are frequently visited in database transactions at least for the user defined number of times which is known as support threshold. Presently a number of algorithms have been proposed in literature to enhance the performance of Apriori Algorithm, for the purpose of determining the frequent pattern.

The main issue for any algorithm is to reduce the processing time. Discovery in Databases (KDD) and Data Mining (DM) helps to extract useful information from raw data. Frequent patterns are those that occur at least a user-given number of times (referred as minimum support threshold) in the dataset. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters. Frequent pattern mining is one of the most important and well researched techniques of data mining. The mining of association rules is one of the most popular research domain. The original motivation for searching association rules came from the need to analyze so called supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. Association rules describe how often items are purchased together. Such rules can be useful for decisions concerning product pricing, promotions, store layout and many others.

3. ASSOCIATION RULE MINING

The association rules mining techniques are applied over databases described as D = {I, T}. Let I = { I_1, I_2, ... ,I_p } be the set of attributes (called items) and T = { t_1, t_2, … t_n } be the transaction set. Each transaction t_i = { I_1, I_2, ... ,I_{mi} } is a set of items, such as $t_i \subset I$ and each subset of items, X, is called itemset [7]. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. [4]

An association rule is an implication $X \rightarrow Y$, where X and Y are two itemsets and $X \cap Y = \phi$. This rule holds on D with the confidence c if c% of transactions in T that contain X, also contain Y. The rule has support s in transaction set T if s% of transactions contain $X \cup Y$.

Since their early definition, association rules are mined using Apriori algorithm proposed for the first time in Agrawal et al., 1993. In association rule mining process, user knowledge can be divided into two main types: domain knowledge, mainly related to database items, and user beliefs expressing user expectations according to the discovered knowledge. In addition, there is a third user-based element described by the actions that a user can realize among his/her different beliefs. Thus, the operators are introduced in order to guide the post-processing step.

Definition 1. Formally, an ontology is a 3-tuple O = {C, R, H}. C = { C_1, C_2, ... ,C_o } is a set of concepts and R = { R_1, R_2, ... ,R_r } is a set of relations defined over concepts. H is a directed acyclic graph (DAG) over concepts defined by the subsumption relation (is-a relation, $\leq$) between concepts. It is said that C_2 is-a C_1, $C_2 \leq C_1$, if the concept C_1 subsumes the concept C_2.

In this approach, there is a domain knowledge model based on ontologies connecting ontology concepts to a set of database items. Consequently, domain ontologies over database extend the notion of Generalized Association Rules based on taxonomies as a result of the generalization of the subsumption relation by the set R of ontology relations. Besides, ontologies are used as filters over items, generating item families.

In this scenario, it is fundamental to connect the ontology to the database, each concept and each instance being instantiated in one/several items.

Considering that the set of concepts C is defined as the union of three concepts subsets $C = C_0 \cup C_1 \cup C_2$:

C_0 is defined as the set of leaf-concepts of the ontology connected in the easiest way to database.

$$C_0 = \{c_0 \in C \mid \nexists c' \in C,\ c' \leq c_0\}$$

In this manner, each concept from C_0 is associated to an item in the database.

$$f_0 : C_0 \to I$$
$$\forall c_0 \in C_0, i \in I, i = f_0(c_0)$$

C_1 is described as the set of generalized concepts in the ontology. A generalized concept is connected to database through its subsumed concepts. That means that, recursively, only the leaf-concepts subsumed by a generalized concept contribute to its database connection.

$$f : C_1 \to 2^I$$
$$\forall c \in C_1, f(c) = \{i = f_0(c_0) \mid c_0 \in C_0, c_0 \leq c\}$$

Based on the concept of strong rules, Rakesh Agrawal et al.[5] introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. More generally, there is the definition of ontology concepts by logical expressions defined over items, organized in the C_2 subset. In a first attempt, there is the base in which there is description of the logical expression on the OR logical operator. Thus the defined concept associated could be connected to a disjunction of items [8][9].

$$f : C_2 \to 2^I, \forall c \in C_2$$
$$c \to E(c)$$
$$f(c) = \{f(c') \mid c' \in E(c)\}$$

To improve association rule selection, there is a rule filtering model, called Rule Schemas. In other words, a rule schema describes, in a rule-like formalism, the user expectations in terms of interesting/obvious rules. As a result, Rule Schemas act as a rule grouping, defining rule families.

The base of Rule Schema formalism is the user representation model introduced by Liu et al. in [6] composed of: General Impressions, Reasonably Precise Concepts and Precise Knowledge. The proposed model is described using elements from an attribute taxonomy allowing an is-a organization of database attributes. A Rule Schema is a semantic extension of the Liu model since it is described using concepts from the domain ontology.

Definition 2. A rule schema is defined as:

$$\langle X_1, X_2, \ldots, X_{s1} \ (\to\ ) Y_1, Y_2, \ldots, Y_{s2} \rangle$$

where X_i and Y_j are ontology concepts and the implication "→" is optional. In other words, it can be noted that the proposed formalism combines General Impressions and Reasonably Precise Concepts. Consequently, if use the formalism as an implication, an implicative rule schema is defined extending the Reasonably Precise Concepts. Meanwhile, if do not keep the implication, there is need to define non implicative rules schemas, generalizing General Impressions [10].

For example, a rule schema $C_2, \overline{C_3} \to C_4$ corresponds to "all association rules whose condition verifies C_2 and doesn't verify the concept C_3, and whom conclusion verifies C_4".

4. PROPOSED WORK

The classical approach should joined with the fuzzy logic controllers for optimization of the results fetched from the data mining and machine learning algorithms. The Association Rule Mining Apriori Algorithm is associated with enormous drawbacks including the iterations involved that reduces the minimum support until it finds the required number of rules with the given minimum confidence. The traditional approach can be improved by overriding some trade-off phases and discarding the unwanted objects and fields from the association analysis. The apriori algorithm needs deep analysis, review as well as revision in terms of the inefficiencies or trade-offs for assorted applications.

- The Classical data mining algorithmic approaches work efficiently on one dimensional database structures.
- Lots of overhead and complexity factors are associated with the classical approach.
- There is the key issue on processing and handling the multi-dimensional database structure and association rule mining.
- The existing systems do not rely and makes use of fuzzy systems for pruning or optimization of the results
- Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all of its proper subsets.
- Lot of other algorithms are under research to identify the maximal frequent item sets without enumerating their subsets, and perform "jumps" in the search space rather than a purely bottom-up approach.
- Apriori Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules.

5. OBJECTIVES OF THE PROPOSED WORK

An Effective Implementation of Rule Based Mining Apriori Algorithm is implemented for multiple applications. The Proposed Approach is applied to the Shopping Cart / Market Basket Analysis Database Domain for fetching the Association Rules and Intelligence.

The implementation is done using following tools :

- WEKA
- MATLAB
- MySQL Database Engine
- MySQL J Connector for JDBC

The Results are analyzed on multiple parameters.

6. TRANSFORMATION OF DATA SET TO RULE MATRIX

1. butter, laptop, curd -> fruit
2. laptop, fruit -> cloth
3. laptop, fruit -> bag
4. laptop, fruit, bag -> bag
5. cloth, laptop, fruit -> bag
6. laptop, fruit -> cloth
7. fruit, bag -> biscuit
8. cloth, milk, grocery -> mouse
9. cloth, mouse, grocery -> tea
10. cloth, mouse, tea -> grocery

The following rules matrix specifies the association rules fetched from WEKA. In the rule matrix, 1's specifies the occurrence of the products in the rules set. 2's refers to the final product that is obtained as the outcome.

$$r = \begin{matrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \end{matrix}$$

```
    0   0   0   1   0   0   0   1   2   0   0
    2   0   0   1   0   0   0   0   0   1   0
    0   1   0   0   0   0   0   1   2   0   0
    0   0   0   0   0   0   0   1   2   0   1
    0   0   0   0   1   0   0   2   0   0
    0   0   1   0   0   0   0   1   2   0   0
    0   0   1   0   0   0   0   0   2   0   0
Total Positive Count in Rule 1
totalr1 =2
Total Positive Count in Rule 1
totalr2 =3
Total Positive Count in Rule 3
totalr3 =2
Total Positive Count in Rule 4
totalr4 =2
Total Positive Count in Rule 1
totalr5 =2
Total Positive Count in Rule 1
totalr6 =2
Total Positive Count in Rule 1
totalr7 =2
Total Positive Count in Rule 1
totalr8 =1
Total Positive Count in Rule 1
totalr9 =2
Total Positive Count in Rule 1
totalr10 =1
totalr =  2    3    2    2    2    2    2    1    2    1
minvalue =
    1
Display All Best Rules
2 3 2 2 2 2 2 2
finalmatrix = 2    3    2    2    2    2    2    2
```

From the final matrix, it is evident that there are 8 rules fetched as optimal results from total 10 rules.

7. CONCLUSION AND FUTURE SCOPE

Association Rule Mining one of the prevalent and decently explored technique for uncovering fascinating relations between variables in expansive databases. It is planned to recognize solid principles ran across in databases utilizing distinctive measures of interestingness Based on the idea of solid standards, Rakesh Agrawal et al. presented the association guidelines for finding regularities between items in substantial scale transaction information recorded by the points of sales frameworks in market basket analysis domain. Such associated data could be utilized as the premise for choices about showcasing exercises, for example, e.g., limited time evaluating or item positions. Notwithstanding the above illustration from business sector bushel dissection affiliation guidelines are utilized today in numerous requisition territories including As far as the future work is concerned, the individual patterns of each object can be analyzed on the web server log files for deep analysis of the links, platform and behavior of the users. For future scope of the work, following techniques can be used in hybrid approach to better and efficient results –

- Particle Swarm Optimization
- HoneyBee Algorithm
- Simulated Annealing
- Genetic Algorithmic Approaches

REFERENCES

[1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases”

[2] Mahendra Pratap Singh Dohare, Premnarayan Arya, Aruna Bajpai, 2012. Novel Web Usage Mining for Web Mining Techniques

[3] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John

Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[4] Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

[5] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925.

[6] Pooja Sharma, Rupali Bhartiya, 2011. An efficient Algorithm for Improved Web Usage Mining, ,Int.J.Computer Technology & Applications,Vol 3 (2), 766-76

[7] Perego, Raffaele, Salvatore Orlando, and P. Palmerini. "Enhancing the apriori algorithm for frequent set counting." In *Data Warehousing and Knowledge Discovery*, pp. 71-82. Springer Berlin Heidelberg, 2001.

[8] Borgelt, Christian, and Rudolf Kruse. "Induction of association rules: Apriori implementation." In *Compstat*, pp. 395-400. Physica-Verlag HD, 2002.

[9] Borgelt, Christian. "Recursion Pruning for the Apriori Algorithm." In *FIMI*. 2004.

[10] Srikant, Ramakrishnan, Quoc Vu, and Rakesh Agrawal. "Mining Association Rules with Item Constraints." In *KDD*, vol. 97, pp. 67-73. 1997.