International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 03 | June-2015 www.irjet.net p-ISSN: 2395-0072

Association Rules Filtration using Dynamic Methods

Monika Mangla[1], Rakhi Akhare[2]

[1] *Asst. Prof., Computer Science Department, LTCE, Navi Mumbai, Maharashtra,* INDIA
[2] *Asst. Prof., Computer Science Department, LTCE, Navi Mumbai, Maharashtra,* INDIA

Abstract - ***The Association rule mining is considered as one of the most relevant tasks in Knowledge Discovery in Databases. Association rule mining increases the discovery of interesting knowledge and valuable information that exists in transaction databases. Due to massive data, huge number of association rules can be discovered, thus it becomes difficult for decision maker to find out the interesting rules and efficient reduction of association rules. To overcome this problem various solutions are proposed by various authors and the best solution for this is elimination of redundant rules from rule base. This paper proposed a new approach for filtering discovered rules by using user and domain expert knowledge. Integration of user knowledge in the post processing is increased by using ontologies connected to data. User expectations are described by the notion of Rule Schema. Rules schemas are used to represent user belief .Combining rule schemas with ontologies forms an interactive framework which improves the interesting rules selection in Post mining process. It saves the time required for Post mining and reduces the redundancies in Association rules.***

Key Words: *Data Exploration, Knowledge Discovery, Ontology, Post mining, Rule Schema.*

1. INTRODUCTION

In this paper we have studied mostly used postmining methods and comparison between different methods. Here we have concentrated on one new interactive post processing approach, ARIPSO (Association Rule Interactive post-Processing using Schemas and Ontology)[2] to eliminate large number of discovered rules discovered during postmining of huge database. Here we have used three concepts ontology, Rule Schema and filters in an interactive framework.. Furthermore, an iterative framework is designed to assist the user throughout the analyzing task. The interactivity of our approach relies on a set of rule mining operators defined over the Rule Schemas in order to describe the actions that the user can perform. As user and domain experts are involved throughout postprocessing, only quality and interesting rules are discovered. As compared to other methods it gives better result as explained in this paper.

This paper is structured as follows. Section 2 describes related work. Section 3 explains definitions used .Section 4 the proposed framework and its elements. Section 5 presents conclusion.

2. DEFINITIONS

Clustering:- Identifying a set of similarity groups in the data.
Classification:- Mining patterns that can classify future data into known classes.
Association Rule:- It is defined as the implication X→Y like if / then statements. Where X and Y are the set of items. Example:- "If a customer buys a dozen eggs, he is 80% likely to purchase milk".
Support :- Measures of how the collection of items in an association occur together as a percentage of all the transactions. Support = # tuples (LHS, RHS)/N
Confidence:-Confidence of rule"X given Y" is a measure of how much more likely it is that Y occurs when A has occurred.
Confidence = # tuples (LHS, RHS) / # tuples (LHS).
An Association rule has two parts:-
Antecedent i.e. if part: - An Antecedent is an item found in data.
Consequent i.e. then part: - A Consequent is an item found in combination with antecedent.
e.g.- Association Rule X→ Y , X – An Antecedent Y – A Consequent.
Frequent Itemset: - An itemset X is called frequent item set in the transaction database D if supp(X) ≥ minsupp.
If X is frequent and no superset of X is frequent, X is denoted as a maximal item set.
Closed Itemset: - A closed itemset is defined as an itemset X which has the property of being the same as its closure, i.e., X= cit(X). The minimal closed itemset containing an itemset Y is obtained by applying the closure operator cit to Y .
Optimal Rule Set*:* - A rule set is optimal with respect to an interestingness metric if it contains all the rules except those with no greater interestingness than one of its more general rules. An optimal rule set is a subset of a nonredundant rule set.

3. LITERATURE REVIEW

Literature review initially starts with Data Mining. There are various data mining algorithms are already exists. Apriori [1] is the first algorithm proposed in the association rule mining

field and many other algorithms were derived from it.

Furthermore, as suggested by Silbershatz and Tuzilin [9], valuable information is often represented by those rare—low support—and unexpected association rules which are surprising to the user.

Different algorithms were introduced to reduce the number of item sets by generating closed [4], maximal [5] or optimal item sets [6], and several algorithms to reduce the number of rules, using non redundant rules [7], [8], or pruning techniques [9].

On the other hand, post processing methods can improve the selection of discovered rules. Different complementary post processing methods may be used, like pruning, summarizing, grouping, or visualization [10].

The CLOSET algorithm was proposed in [11] as a new efficient method for mining closed item sets. CLOSET uses a novel frequent pattern tree (FP-tree) structure, which is a compressed representation of all the transactions in the database. Moreover, it uses a recursive divide-and-conquer and database projection approach to mine long patterns. Another solution for the reduction of the number of frequent item sets is mining maximal frequent item sets [5].

Bayardo, Jr., et al. [12] proposed a new pruning measure (Minimum Improvement) described as the difference between the confidences of two rules in a specification/generalization relationship. The specific rule is pruned if the proposed measure is less than a pre-specified threshold, so the rule does not bring more information compared to the general one. The Rule Schema formalism is based on the specification language for user knowledge introduced by Liu et al. [5].

The first idea of using Domain Ontologies was introduced by Srikant and Agrawal with the concept of Generalized Association Rules (GAR) [13]. The authors proposed taxonomies of mined data (an is-a hierarchy) in order to generalize/specify rules.

Another contribution, very close to [13], uses taxonomies to generalize and prune association rules. The authors developed an algorithm, called GART , which, having several taxonomies over attributes, uses iteratively each taxonomy in order to generalize rules, and then, prunes redundant rules at each step. The item-relatedness filter was proposed by Natarajan and Shekar [8].

Limitations:

- The number of frequent closed item sets generated is reduced in comparison with the number of frequent item sets.
- The huge number of discovered rules makes very difficult for a decision maker to manually outline the interesting rules. Thus, it is crucial to help the decision maker with an efficient reduction of the number of rules.
- It is crucial to help the decision-maker with an efficient technique for reducing the number of rules.

4. THE INTERATIVE FRAMEWORK

The new approach i.e. ARIPSO defines a new formal environment to prune and group discovered associations integrating knowledge into specific mining process of association rules. It is composed of three main parts (as shown in Figure).Firstly, a basic mining process is applied over data extracting a set of association rules. Secondly, the knowledge base allows formalizing user knowledge and goals. Domain knowledge allows a general view over user knowledge in database domain, and user expectations express user already knowledge over the discovered rules. Finally, the post-processing step consists in applying several operators (i.e. pruning) over user expectations in order to extract the interesting rules.The novelty of this approach is user knowledge representation in the form of one or several ontologies and several rule schemas.

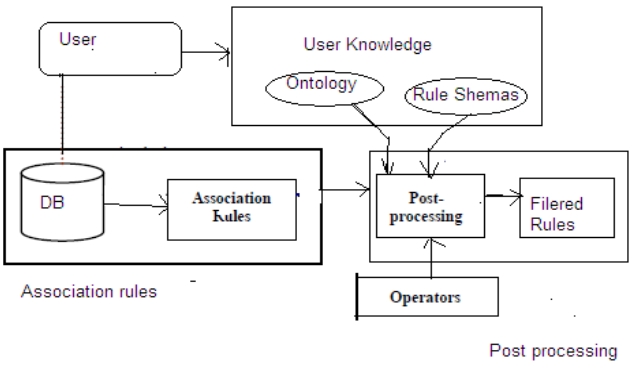

Fig -1: Interactive Framework

As shown in above Fig. framework divided into basically two parts first is knowledge base involves user knowledge and goals and second is post processing task. Domain knowledge offers a view over user's knowledge about database domain. In second part set of filters are applied over extracted rules to select Interesting rules.

ARIPSO framework involves six different steps for selecting interesting rules:

***1. Ontology Construction*:-** Ontology defines as explicit specification of shared conceptualization i.e. abstract model of some phenomenon.One of our most important contributions relies on using ontologies as user background knowledge representation. Thus, we extend the specification language General Impressions (GI), Reasonably Precise Concepts (RPC), and Precise Knowledge (PK)—by the use of ontology concepts.

In this approach, we propose 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.

2. Rule Schema Definition:- To improve association rule selection, we propose 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. Rule schema is the way of representation of ontologies. It is simplification of concepts used in ontology defined.

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. We propose to develop two of the three representations introduced in [13]: General Impressions and Reasonably Precise Concepts. Thus, rule schemas bring the complexity of ontologies in rule mining combining not only item constraints, but also ontology concept constraints.

For example, a rule schema *C1, ~C2 → C3* corresponds to "all association rules whose condition verifies *C1* and doesn't verify the concept *C2*, and whom conclusion verifies *C3*".
Rule Schema combines concepts of General Impressions and Reasonably Precise Concepts.

3. *Selection of Operators*:- There are four operators applied over rule schemas to eliminate uninteresting rules discovered during postprocessing. These operators are Pruning operator, Conforming operator, Unexpectedness operator and Exception operator

The Pruning Operator removes all association rules matching to rule schema. For this matching conforming operator is used. It is denoted as P(RS).

The conforming *Operator* applied over a rule schema, *C(RS)*, used to find the implication between several concepts. Result of this is, rules matching all the elements of a non-implicative rule schema are filtered. For an implicative rule schema, the condition and the conclusion of the association rule should match those of the schema.

The *unexpectedness operator, U(RS)*, used to filter a set of rules which surprises to for the user. AS user mostly searches to discover new knowledge by using its his/her prior knowledge, this operator preferably used by user than conforming operator.

several types of unexpected rules can be filtered according to the rule schema: rules unexpected regarding the antecedent Up, rules unexpected regarding the consequent Uc, and rules unexpected regarding both sides Ub.

The *exception operator* applied over *RS1*, is defined only over implicative rule schemas and extracts conforming rules with respect to the following new implicative rule schema: X ^ Z →Y where Z is a set of items.

4. *Visualizing and Validation*:-Pruned association rules are proposed to user for validating the results and revise his/her knowledge and expectations.

5. *Filters*:- To reduce the number of rules two types of filters are applied over selected rules. These filters are as below:

- ***MICF (Minimum Improvement Constraint Filter):-***It is used to filter those rules whose confidence is greater than that of its any other simplification.
  Bread, Eggs → milk (Confidence = 75%);
  Bread → milk (Confidence= 95%);
  Eggs → milk (Confidence = 64%):

Here, last two rules are simplification of first rule. As confidence of first rule is not greater than its simplified rule, it is not selected.

- ***Item-Relatedness Filter (IRF)***:- It was proposed by Shekar and Natarajan. Item relatedness means semantic distance between items in item taxonomies. users are interested to find association between itemsets with different functionalities, coming from different domains.

The distance between each pair of items from the condition and, respectively, the consequent is computed as the minimum path that connects the two items in the ontology. Thus, the item-relatedness (IR) for a rule is defined as the minimum of all the distance computed between the items in the condition and the consequent: e.g. R1: Bread, Eggs, Apple→ milk

IR(R1)=min(d(Bread,milk),d(Eggs,Milk),d(Apple,milk))
=min (5, 5, 3) =3

As shown in above example minimum distance rule i.e. R (Apple, Milk) is filtered out.

6. *Iterative Interaction*: - interactive approach applied iteratively in post processing helps user to revise his/her knowledge proposed by them. By using this type of loop user can modify rule schema, change the operator. Again he/she can decide type of filers wants to apply over rules.

In this framework, user knowledge is integrated into association rule mining using ontologies and rule schemas. This framework proposed is successful to deliver significant rules but ontologies and rule schema design is not possible without Domain Expertization and format of rule schemas.

As Apriori algorithm is expensive, Closet algorithm requires special type of data structure i.e. FP Tree, GART requires several type of taxonomies, time required is more and in M-SQL there is lack of user knowledge exploration. As compared with these methods, our framework is simple and useful for user because users is involved and revise his/her knowledge throughout the postprocessing steps. Domain Expert revises the quality of rules so it gives better results.

5. CONCLUSIONS

By applying this new interactive framework over a large database, we allowed the integration of user knowledge in the post processing steps in order to reduce the number of rules to several dozens or less. During interactive process the quality of the filtered rules is validated by the domain expert. Thus, by using this framework, user can select interesting association rules throughout huge volumes of discovered rules. It saves the time and efforts in post mining by eliminating the useless, redundant rules. It gives the quality results as knowledgeable people involved throughout the process. The more the knowledge is represented in a flexible expressive, and accurate formalism, the more the rule selection is efficient. It will be very useful for the user to be able to introduce in the GI language interesting additional information. The representation of user expectations is more general, and thus, filtered rules are more interesting for the user. As user is involved throughout post mining task, it is possible to select interesting rules.

REFERENCES

[1] Text Book "Fundamentals of Database Systems"- by R.Elmasari and S.B.Navathe
[2] Post-Processing of Discovered Association Rules Using Ontologies"- Claudia Marinica, Fabrice Guillet and Henri Briand
[3] Interactive Post mining of Association Rules by Validating Ontologies"-ijecse ISSN- 2277-1956
[4] Survey on Post mining Methodologies of Association rules"by KalliSrinivasaNageswara Ramakrishna,Volume 3, May2011.
[5] B. Liu, W. Hsu, L.-F. Mun, and H.-Y. Lee, "Finding Interesting Patterns Using User Expectations," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 817-832, Nov. 1999.
[6] "Interactive Approach for Postmining of Asspciation Rules" Claudia Marinica, Fabrice Guillet,vol.22,No.6,June 2010.
[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, pp. 207-216, 1993.
[8] R. Natarajan and B. Shekar, "A Relatedness-Based Data-Driven Approach to Determination of Interestingness of Rules," Proc. 2005 ACM Symp. Applied Computing (SAC), pp. 551-552, 2005..
[9] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interestingin Knowledge Discovery Systems," IEEE Trans. Knowledge and Data Eng. vol. 8, no. 6, pp. 970-974, Dec. 1996.
[10] J.Pei,J.Han and R.Mao,"Closet:An Efficient Algorithm for Mining Frequent Closed Itemsets",Proc.ACM SIGMOD Workshop Research,pp.21-30,2000
[11] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," Proc. Seventh Int'l Conf. Database Theory (ICDT '99), pp. 398-416, 1999.
[12]]R.J.Bayardo,Jr.,R.Agrawal,"Constraint based Rule Mining in Large, Dense Databases".
[13] R.Shrikant and R.Agrawal,"Mining Generalized Association Rules" Proc.21st Int conf.Very Large Database,pp407-419,1995.

