International Journal of Computer and Communication Technology

Volume 6 | Issue 3 Article 10

July 2015

POST-MINING OF ASSOCIATION RULES USING ONTOLOGIES AND RULE SCHEMAS

R. SUBASH CHADRA BOSE
Computer Science, A.V.V.M.Sri Pushpam College, Poondi-613 503, Thanjavur-Dt, rsbose.cs@gmail.com

R. SIVAKUMAR
Computer Science, A.V.V.M.Sri Pushpam College, Poondi-613 503, Thanjavur-Dt, rskumar.avvmspc@gmail.com

Recommended Citation

BOSE, R. SUBASH CHADRA and SIVAKUMAR, R. (2015) "POST-MINING OF ASSOCIATION RULES USING ONTOLOGIES AND RULE SCHEMAS," *International Journal of Computer and Communication Technology*: Vol. 6 : Iss. 3 , Article 10.
DOI: 10.47893/IJCCT.2015.1303
Available at: https://www.interscience.in/ijcct/vol6/iss3/10

POST-MINING OF ASSOCIATION RULES USING ONTOLOGIES AND RULE SCHEMAS

R. SUBASHCHADRABOSE[1] & R.SIVAKUMAR[2]

[1,2]Computer Science, A.V.V.M.Sri Pushpam College, Poondi-613 503, Thanjavur-Dt
Email:rsbose.cs@gmail.com, rskumar.avvmspc@gmail.com

Abstract-Knowledge discovery and databases (KDD) deals with the overall process of discovering useful knowledge from data. Data mining is a particular step in this process by applying specific algorithms for extracting hidden fact in the data. Association rule mining is one of the data mining techniques that generate a large number of rules. Several methods have been proposed in the literature to filter and prune the discovered rules to obtain only interesting rules in order to help the decision-maker in a business process. We propose a new approach to integrate user knowledge using ontologies and rule schemas at the stage of post-mining of association rules.

General Terms- Lattice, Post-processing, pruning, itemset

Keywords-*Support, confidence, unexpectedness, actionability*

1. INTRODUCTION

1.1 Association Rule

The association rule mining is one of important mining tasks in KDD [10]. It can be stated as follows: Let I={1,2,3,.....,n} be a set of items ,and let T={1,2,3,....,n} be a set of transaction identifier or tids. The input database is a binary relation of $\delta \subseteq I \times T$. If an item occurs in a transaction t, then it is denoted as $(i,t) \in \delta$ or as $i\ \delta\ t$. The support of an itemset X is denoted by σ (x) that is the number of transaction in which it occurs as a subset. An itemset is frequent if its support is more than or equal to user specified minimum support (minsup) threshold that is $\sigma(x) \geq$ minsup. An association is an expression $A \xrightarrow{p} B$, where A and B are itemsets and $A \cap B = \varphi$. The support of a rule is given as $\sigma(AUB)$ that is the joint probability of a transaction containing both A and B and the confidence is denoted by $p= \sigma(AUB)/ \sigma(A)$ that is the conditional probability than a transaction contains B, given that it also is contains. A rule is frequent if the itemset AUB is frequent. A rule is confident if its confidence is greater than or equal to a user-specified minimum confidence (minconf), threshold value that is $p \geq$ minconf.

1.2 Post-Mining

Association rule mining produces a large amount of rules. In the literature, several algorithms have been proposed to discover and maintain association rules. However, one of the problems of association rules still remain unsolved is the number of discovered rules will be huge. To overcome this drawback, the post-processing task has been proposed to improve the selection of discovered rule. Different complementary post-processing methods like pruning, summarizing, grouping and visualization [4] may be used to empower the selection of association rule. Clearly, the large number of rules make difficult to imagine by a user. Moreover, many of the discovered rules will be uninteresting, obvious or irrelevant. For this reason, there are so many methods have been proposed to assist a user in detecting the most interesting and relevant once. Measures of rule interestingness can be divided into two categories such as subjective and objective. Subjective measure of interestingness focuses on finding interesting patterns by matching against a given set of user beliefs. Objective measure of interestingness is in terms of their statistical significances. In subjective analysis, prior domain knowledge is used to determine unexpectedness in the extracted association rules. The interestingness of an extracted association rule can be described in terms of its unexpectedness and actionability. Expected rule confirm prior domain knowledge and are essentially known, unexpected rules are novel rules which are previously unknown which may contradict the user's existing knowledge. A rule is actionable if a domain expert can use it to his/her advantage. Since prior domain knowledge is required to determine unexpectedness in terms of which rules are known or novel. To reduce the effort required to identify interesting rules, researchers have offered approaches that aim to minimize the number of rules generated.

This paper proposes a new approach to strengthen the integration of user knowledge using domain ontologies and rule schemas. Furthermore, a framework is designed to assist the decision-maker for association rule analysis task. User expectations are represented by rule schemas and rule operators are used to guide user actions. Ontologies will provide a mechanism to represent user knowledge.

The paper is structured as follows. Section 2 gives the details of research area and reviews of related works. Section 3 describes the framework design. Section 4 describes the ontology and rule schema. At last, section 5 presents conclusion and point to the directions of future research.

2. RELATED WORKS

In general, the most appropriate method of analyzing very large binary data set is association rule analysis. An application of this analysis is to discover relationships between binary variables in transactional databases. If association rules used perhaps to analyze non-binary data, the data being coded as binary data. There are two or three steps involved in association rule analysis: 1. coding of data as binary if data is not binary 2. rule generation and 3. post-mining. This paper focuses on the third step. Rule generation was first introduced by Agrawal et al. in the year 1993, since then most of the published works were focused on effective and fast techniques for rule generation. Mining was performed by generating rules with support and confidence above the user-defined thresholds. The generation of association rules can be divided into two steps: 1. finding frequent itemsets and 2. generating high-confidence rules. A frequent itemset is an itemset with support above the user-defined threshold. Several research works have been focused on efficient searching strategy for finding frequent itemset through the large number of candidate itemsets to improve the efficiency of searching; it exploits the downward-closure or anti-montonicity property of support: if an itemset is frequent if all of its subsets are frequent. This property is empowered to traverse the search space efficiently. The apriori algorithm is one of the most commonly used methods for discovering association rules. Although, apriori has been applied successfully in many cases, it does have performance problems, and so several algorithms have been proposed to improve apriori's performance. The threshold values such as minimum support and minimum confidence count have to be supplied by the domain expert to the apriori algorithm and in subsequent improvements made by trial and error method. If threshold values are set to high, only a small number of rules will be generated. If they are set to low, too many rules will be generated. The most important and problematic steps in an association rule discovery process is the post-processing of the extracted association rules that is the interpretation, evaluation and validation. The number of association rules generated by the apriori algorithm and the subsequent improvement of this algorithm can be huge and thus also making manual analysis of the rule is difficult. Alternative algorithms were proposed in the literature to reduce the number of candidate itemset by discovering closed itemset or maximal frequent itemset. A closed itemset is an itemset without any superset having the same support. The ECLAT algorithm was introduced by Zaki et al. is a viable alternative to the apriori algorithm that is based on equivalence class clustering with some support threshold.

There are two different approaches for searching itemset lattice such as breadth-first search and depth-first search. Aprori is based on a breadth-first search. Depth-first search of itemset lattice was proposed by Burdick et al. in the year 2001. Han et al. designed a frequent-pattern tree (FP-tree) as a compressed representation of database in the year 2004 and introduced the FP-growth algorithm for mining the complete set of frequent patterns. Throughout the literature, there have been some alternative to support based rule generation like conditional independencies. The algorithm MAMBO designed by Castelo et al. (2001) based on conditional independencies. Pasquier et al. proved that it is sufficient to mine the closed item frequent set not supposed to used support threshold. Pie et al. (2000) and Zaki and Hsiao(2005) where proposed algorithm CLOSET and CHARM respectively for this purpose.

The major challenge of data mining is not only to improve its efficiency but also how to improve its interpretability of discovered rules. During mining process, a huge number of rules discovered when the real-world database is large to be explored. Furthermore, maintaining of these rules turned out to be extremely costly and difficult. So, users can't interpret and understand those overwhelming number of rules. Hence, it is an urgent need for intelligent techniques to handle useless rules and help users to understand the results from the rapidly growing volumes of digital data. The purpose of post-processing is to enhance the quality of the mined knowledge. It can help end-users to understand and interprets the meaning of knowledge nuggets.

In the field of data mining research, many attentions have been paid in dealing with more knowledge through measuring similarity or redundancy. To measure the similarity between rules, Eucledean distance (Waitman et al.,) is used to discard the rules with high similarity. In addition, chi-square test(Liu et al.,1999) and entropy (Jarozewicz and Simovici 2002) are used to analyze the distance between the rules in the post-processing step. The issue of rules importance (Geng and Hamilton, 2006) has been studied by considering both measures of objective and subjective. For example Brin et al. (1997) outlined a conviction measurement to express rule interestingness. The typical case to prune the generated rule base directly is rule cover technique proposed by Toivonen et al. (1995) where a rule is redundant if it is covered by others . Its performance is improved by a technique called integer programming devised by Brijs et al. (2000).

However, Klemettinen et al. (1994)extracted interesting rules from the entire rule base through rule templates[12] whereas Srikant et al. (1995) organized a rule base as a hierarchical structure and the rules are more general in higher level than those are in low level. Liu et al. (1999) divided a rule base into two parts by means of direction setting (DS) and Non-direction setting(non-DS). Under this scheme DS part have key information while those non-DS contains relevant detailed knowledge .Recently, Ashrafi et al. (2007) proposed fixed antecedent and consequent methods to discard redundant rules from the resultant rule set. In this method, a rule is redundant when a set of rules that also convey the same knowledge is found

The above mentioned methods are not only based on rule structure or interestingness measurements. This may of cause a problem that redundant or insignificant rules still exist in the pruned rules. In this regards, a fast rule reduction algorithm using closed itemset was introduced. This algorithm was implemented in two steps. The first step is performing data mining algorithms on database and the second step is to carry out closed mining method on the mined rule base to obtain closed rule subsets.

In the literature review, we noticed that many number of proposals that exist to address the problem of redundant rule elimination. These proposals can be classified into two categories: pre-pruning and post-pruning. The task of pre-pruning is to prevent redundant rules at the mining stage, which is the pruning operation, occurs during the phase of rule generation, possibly the generated rules are all significant. By contrast post-pruning technique concerns that the pruning operation occurs often rules have been generated. Since it occurs at the post-processing step and is independent of mining algorithms. This approach has many outstanding attracted pruning methods that have been developed. However, the main disadvantage of these approaches is that the computational cost is very high when the number of association rules is huge. This method exploited approximate functional dependency between rules to eliminate superfluous rules. Moreover, it generated only closed rule-sets which is far less than association rules and discarded redundancy between discovered rules.

In the literature, several methods proposed to discard redundant or useless rules based on the following concepts are shown in Table 1.

Table 1 Survey of post-mining methods

Author	Year	Concept
Toivennen et al.	1995	Rule cover
Balalis and Psaila	1997	Template based constraint
Fernadez	2001	Rough sets
Baralisi and Chiusano	2004	Essential rule set
DoninnQagues and Rezende	2005	Knowledge taxonomy
Bathoorn et al.	2006	Minimum description length principles

3. FRAMEWORK DESCRIPTION

The new approach defines an environment to integrate user knowledge which is represented through ontology and rule schema. It consists of three main parts is shown in figure 1. First, the basic mining method is applied to mine association rules. Second, the knowledge base allows formalizing user knowledge. Domain knowledge gives a generic view over user knowledge in database domain and applying a set of filters iteratively over discovered rules. Finally, the post- processing part consists of several operators which are applied over user expectations to filter interesting rules. The novelty of this approach envisages in supervising the knowledge discovery process with different conceptual structures for user knowledge representation of one or more ontologies and rule schemas.

In association rule mining, user knowledge can be divided into two main parts such as domain knowledge, related to database items, and user beliefs expressing user expectations according to the discovered knowledge. We introduce a set of operators to guide the post-processing step by the actions that a user can realize his/her beliefs. The rule schema filter is based on operators over rule schemas. This allows users to perform several actions over discovered rules. It uses four kinds of operator: conformity, unexpectedness, pruning and exceptions. Filters are used to reduce the number of rules. Three filters are included in the framework: rules schema filter, minimum improvement constraint and item-relatedness filter

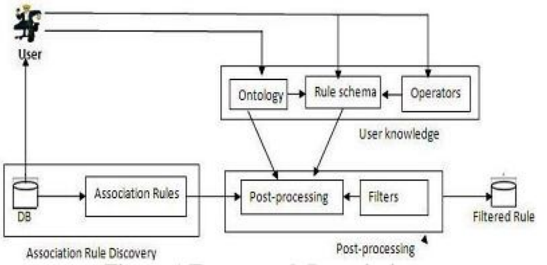

Figure 1 Framework Description

4. ONTOLOGY DESCRIPTION

In 1990s, ontology [11] was proposed by Gruber as formal, explicit specification of shared conceptualization. By conceptualization it is meant for abstract model of some phenomenon described in terms of concept. The formal notion denotes the idea that machines should be able to interpret ontology. Explicit refers to the transparent definition of ontology elements. Finally, shared outlines ontology that brings together some knowledge common to a certain group. In the literature, four types of ontologies have been proposed: upper (top-level) ontologies, domain ontologies, task ontologies and application ontologies. Top-level ontologies deal with general concept and the other types deal with domain specific concept. Ontologies were introduced for the first time in the early 2000s. It can be used in several ways: Domain and Background Knowledge ontologies organize domain knowledge and play important roles at several levels of the knowledge discovery process, ontologies for data mining process codify mining process description and choose the most appropriate task according to the given problem and metadata ontologies described the construction process of the items.

In this paper, we focus on Domain and Background Knowledge ontologies. The specification language proposed by Liu et al. [13] proposed an approach to represent user expectations[17][15] with the discovered rules using three levels of specification: General Impression(GI), Reasonably Precise Concepts(RPC) and Precise Knowledge(PK). The authors developed a representation formalism which is very close to the association rule formalism, flexible enough and comprehensible for the user. The syntax of the GI is

gi(<S1,….Sm>)[Support,Confidence]

where Si is an element of an item taxonomy or an expression defined using *+/? operators, and support and confidence thresholds.

In addition, it is difficult for a domain experts to know exactly the support and confidence threshold for each rule schema proposed, because of their statistical significance so that we use PK in user expectation representation is useless. Thus, the other two representations: GI & RPC are introduced in [13].

A rule schema expresses the fact that the user expects certain elements to be associated in the extracted association rules. This can be expressed as.

RS(<X1,……..Xn(→)Y1…..Ym>),

Where Xi,Yj Є C of O={C,R,I,H,A} and the implication '→' is optional. In other words, we can note that the proposed formalism combines General Impressions and Reasonably Precise Concepts. If we use the formalism as an implication, an implicative Rule Schema is defined extending the RPC. On the other hand, if we do not keep the implication, we define nonimplicative Rules Schemas, generalizing GI.

It is fundamental to connect ontology concepts C of O={C,R,I,H,A} to the database, each one of them being connected to one/several items of I. To this end, we consider three types of concepts: leaf-concepts, generalized concepts from the subsumption relation (≤) in H of O, and restriction concepts proposed only by ontologies. In order to proceed with the definition of each type of concepts, let us remind that a set of items in a database is

defined as I={i_1, i_2,….i_n}.

The leaf-concepts (C0) are defined as

C0={c0 Є C |∃c' ∃Є C, c' ≤ c0}.

Each concept from C0 is associated to one item in the database:

f0 : c0→I, ∀c0 Є C0, ∃i Є I, i= f0(c0).

Generalized concepts (C1) are described as the concepts that subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts. This means that, recursively, only the leaf-concepts subsumed by the generalized concept contribute to its database connection:

$$f : C_1 \to 2^I$$
$$\forall c_1 \in C_1, f(c_1) = \bigcup_{c_0 \in C_0} \{i = f_0(c_0) \mid c_0 \leq c_1\}.$$

Restriction concepts are described using logical expressions defined over items and are organized in the C2 subset. In a first attempt, we use the description of the concepts on restrictions over properties available in description logics. Thus, the restriction concept defined could be connected to a disjunction of items.

5. CONCLUSION

This paper discusses the problem of helping the decision-maker, to provide the interesting patterns in the post-processing stage of association rule mining. Rule schemas allow user knowledge representation together with ontologies to improve the selection of interesting rule. Our aim is to improve the approach by the way of developing the rule schema formalism and integrating this approach in the knowledge discovery algorithm to obtain useful rules.

REFERENCES

[1] Adomavicius, G. and A. Tuzhilin. (2001). Expert-Driven Validation of Rule-Based User Models in Personalization Applications. Data Mining and Knowledge Discovery, pages 33–58.

[2] Agrawal, R., T. Imielinski, and A. Swami. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 12th ACM SIGMODInternational Conference on Management of Data, pages 207 - 216.

[3] An, A., S. Khan, and X. Huang. (2003). Objective and Subjective Algorithms for Grouping Association Rules. International Conference in Data Mining, pages 477 - 480.

[4] Baesens, B., S. Viaene, and J. Vanthienen. (2000). Post-Processing of Association Rules. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000), pages 2 - 8.

[5] Berrado, A. and G. C. Runger. (2007). Using metarules to organize and group discovered association rules. Data Mining and Knowledge Discovery, pages 409 – 431.

[6] Bayardo, R.J. Jr. and R. Agrawal. (1999). Mining the most interesting rules. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 145 – 154.

[7] Češpivová, H., J. Rauch, V. Svátek, M. Kejkula, and M. Tomečková. (2004). Roles of Medical Ontology in

Association Mining CRISP-DM Cycle. Workshop Knowledge Discovery and Ontologies in ECML/PKDD

[8] Chawla, S., J. Davis, and G. Pandey. (2004). On Local Pruning of Association Rules Using Directed Hypergraphs. Proceedings of the 20th International Conference on Data Engineering, pages 832.

[9] Euler, T. and M. Scholz. (2004) Using Ontologies in a KDD Workbench. In Workshop on Knowledge Discovery and Ontologies at ECML/PKDD.

[10] Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. (1996) Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press.

[11] Gruber, T. (1993). A translation approach to portable ontology specification. Knowledge Acquisition, 5:199- 220.

[12] Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. (1994). Finding Interesting Rules from Large Sets of Discovered Association Rules. International Conference on Information and Knowledge Management (CIKM), pages 401-407.

[13] Liu, B., W. Hsu, K. Wang and S. Chen. (1999). Visually Aided Exploration of Interesting Association Rules. Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining, Lecture Notes In Computer Science, Vol. 1574, Springer-Verlag, pages 26 – 28.

[14] Nigro, H.O., S.E. Gonzalez Cisaro, and D.H. Xodo. (2007) Data Mining With Ontologies: Implementations, Findings and Frameworks, Idea Group Reference

[15] Padmanabhan, B. and A. Tuzhuilin. (1999). Unexpectedness as a Measure of Interestingness in Knowledge Discovery. Decision Support Systems,Volume 27, Number 3, Elsevier, pages 303-318.

[16] Piatetsky-Shapiro, G. and C.J. Matheus. (1994). The Interestingness of Deviations. In U. M. Fayyad and R. Uthurusamy (eds.), Knowledge Discovery in Databases,Papers from AAAI Workshop (KDD ' 94), pages 25 – 36.

[17] Silberschatz, A. and A. Tuzhilin. (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Transactions on Knowledge and Data Engineering, IEEE Educational Activities Department, pages 970 - 974.

[18] Srikant, R. and R. Agrawal. (1995). Mining Generalized Association Rules. In U. Dayal, P.M.D. Gray, and S. Nishio, eds, Proceedings of the 21st International Conference on Very Large Databases, pages 407 – 419.

[19] Toivonen, H., M. Klementinen, P. Ronkainen, K. Hatonen, and H. Mannila. (1995). Pruning and Grouping of Discovered Association Rules. Mlnet Workshop on Statistics, Machine Learning, and Discovery in Databases, pages 47 - 52.

[20] Zhou, X. and J. Geller. (2007). Raising, to enhance rule mining in web marketing with the use of an ontology. Date Mining with Ontologies: Implementations, Findingsand Frameworks, pages 18-36.

[21] Claudia Marinica and Farice Guilet. (2010). Knowledge-Based Interactive Postmining of Association Rules using Ontologies. IEEE Transactions on Knowledge and Data Engineering, pages 784 – 797.

◈◈◈