International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 7029-7032

Rank-Based Weighted Rule Mining Using Post Mining Methods with Ontology Support

G. Silambarasan
Research Scholar, CMJ University, Meghalaya, India.
Orcid Id: 0000-0002-8342-0128

Dr. T. Anand
Professor/ Head, Dept. of Computer Science and Engineering, Vels University, Chennai, TamilNadu, India.
Orcid Id: 0000-0002-4034-7143

Dr. V. Chandrasekar
Associate Professor, Dept. of Computer Science and Engineering,
Malla Reddy College of Engineering and Technology, Secunderabad, Telangana State, India.

Orcid Id: 0000-0001-7258-0794

Abstract

Association rule mining is used to extract frequent item sets. Apriori algorithm is used to mine association rules. Minimum support and confidence values are used to identify interested rules. Low support threshold produces more number of rules. The rules are used for the decision making process. Rule reduction is required for efficient decision-making system. Knowledge based rule reduction schemes are used to filter the interested rules.

Post mining schemes are used to filter derived rules. Pruning, summarizing, grouping, and visualization techniques are used for the post mining process. Uninterest or redundant rules are removed in pruning process. Concise sets of rules are generated in summarizing method. Grouping process produces groups of rules. Visualization produces graphical format of rules. Association Rule Interactive post-Processing using Schemas and Ontologies (ARIPSO) mechanism is used for post mining process. ARIPSO is used to prune and filter discovered rules. In the existing system rule validation is not provided. Quantitative attributes are not considered in the post-mining scheme. Weighted rule mining scheme is not supported.The proposed system is designed to perform post mining on derived rules with ontology support. The rule-mining scheme is enhanced to handle quantitative attributes. ARIPSO scheme is enhanced with validation methods. Weighted rule mining and filtering process can be integrated with the ARIPSO scheme. Rank based concept relationship analysis can be provided to improve the post mining process.

Keywords: Ontology, Association rules, ARIPSO mechanism, Weighted rule mining.

INTRODUCTION

Association rule mining, introduced in is considered as one of the most important tasks in Knowledge Discovery in Databases. Among sets of items in transaction databases, it aims at discovering implicative tendencies that can be valuable information for the decision-maker. An association rule is defined as the implication X! Y,

Described by two interestingness measures—support and confidence—where X and Y are the sets of items. Apriority is the first algorithm proposed in the association rule mining field and many other algorithms were derived from it. Starting from a database, it proposes to extract all association rules satisfying minimum thresholds of support and confidence. It is very well known that mining algorithms can discover a prohibitive amount of association rules; for instance, thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions. Furthermore, as suggested by Silbershatz and Tuzilin , valuable information is often represented by those rare—low support—and unexpected association rules which are surprising to the user. So, the more we increase the support threshold, the more efficient the algorithms are and the more the discovered rules are obvious, and hence, the less they are interesting for the user. As a result, it is necessary to bring the support threshold low enough in order to extract valuable information. Unfortunately, the lower the support is, the larger the volume of rules becomes, making it intractable for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules overpasses 100. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. To overcome this drawback, several methods

were proposed in the literature. On the one hand, different algorithms were introduced to reduce the number of itemsets by generating closed , maximal or optimal itemsets , and several algorithms to reduce the number of rules, using nonredundant rules or pruning techniques . On the other hand, postprocessing methods can improve the selection of discovered rules. Different complementary postprocessing methods may be used, like pruning, summarizing, grouping, or visualization Pruning consists in removing uninteresting or redundant rules. In summarizing, concise sets of rules are generated. Groups of rules are produced in the grouping process; and the visualization improves the readability of a large number of rules by using adapted graphical representations.

RELATED WORK

Concise Representations of Frequent Itemsets

Interestingness measures represent metrics in the process of capturing dependencies and implications between database items, and express the strength of the pattern association. Since frequent itemset generation is considered as an expensive operation, mining frequent closed item sets was proposed in order to reduce the number of frequent itemsets. For example, an itemset X is denoted as closed frequent itemset i. Thus, the number of frequent closed itemsets generated is reduced in comparison with the number of frequent itemsets. The CLOSET algorithm was proposed in a new efficient method for mining closed itemsets. 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.

Redundancy Reduction of Association Rules

Conversely, generating all association rules that satisfy the confidence threshold is a combinatorial problem. Zaki and Hsiao used frequent closed itemsets in the CHARM algorithm in order to generate all frequent closed itemsets. They used an itemset-tid set search tree and pursued with the aim of generating a small nonredundant rule set. To this goal, the authors first found minimal generator for closed itemsets, and then, they generated nonredundant association rules using two closed itemsets. Acquire proposed the Close algorithm in order to extract association rules. Close algorithm is based on a new mining method: pruning of the closed set lattice (closed itemset lattice) in order to extract frequent closed itemsets. Association rules are generated starting from frequent itemsets generated from frequent closed itemsets. More recently, Li proposed optimal rules sets, defined with respect to an interestingness metric. An optimal rule set contains all rules except those with no greater interestingness than one of its more general rules.

A set of reduction techniques for redundant rules was proposed and implemented. The developed techniques are based on the generalization/specification of the antecedent/consequent of the rules and they are divided in methods for multiantecedent rules and multiconsequent rules. Hahsler et al. were interested in the idea of generating association rules from arbitrary sets of itemsets. This makes possible for a user to propose a set of itemsets and to integrate another set generated by a data mining tool. In order to generate rules, a support counter is needed; consequently, the authors proposed an adequate data structure which provides fast access: prefix trees.

User-Driven Association Rule Mining

Interestingness measures were proposed in order to discover only those association rules that are interesting according to these measures. They have been divided into objective measures and subjective measures. Objective measures depend only on data structure. Many survey papers summarize and compare the objective measure definitions and properties. Unfortunately, being restricted to data evaluation, the objective measures are not sufficient to reduce the number of extracted rules and to capture the interesting ones. Several approaches integrating user knowledge have been proposed. In addition, subjective measures were proposed to integrate explicitly the decision-maker knowledge and to offer a better selection of interesting association rules. Silbershatz and Tuzilin proposed a classification of subjective measures in unexpectedness—a pattern is interesting if it is surprising to the user—and actionability—a pattern is interesting if it can help the user take some actions. As early as 1994, in the KEFIR system, the key finding and deviation notions were suggested. Grouped in findings, deviations represent the difference between the actual and the expected values. KEFIR defines interestingness of a key finding in terms of the estimated benefits, and potential savings of taking corrective actions that restore the deviation back to its expected value. These corrective actions are specified in advance by the domain expert for various classes of deviations. Later, Klemettinen proposed templates to describe the form of interesting rules (inclusive templates) and not interesting rules (restrictive templates). The idea of using templates for association rule extraction was reused. Other approaches proposed to use a rule-like formalism to express user expectations, and the discovered association rules are pruned/summarized by comparing them to user expectations.

Ontologies in Data Mining

Depending on the granularity, four types of ontologies are proposed in the literature: upper (or top level) ontologies, domain ontologies, task ontologies, and application ontologies. Top-level ontologies deal with general concepts; while the other three types deal with domainspecific concepts.

Ontologies, introduced in data mining for the first time in early 2000, can be used in several ways: Domain and Background Knowledge Ontologies, Ontologies for Data Mining Process, or Metadata Ontologies. 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; while Metadata Ontologies describe the construction process of items. In this paper, we focus on Domain and Background Knowledge Ontologies. The first idea of using Domain Ontologies was introduced by Srikant and Agrawal with the concept of Generalized Association Rules (GAR). The authors proposed taxonomies of mined data in order to generalize/specify rules.

DATA PREPROCESS

The data preprocess module is used to normalize the data values Customer survey information data set is used in the system. Noise elimination process is performed to reduce redundant data values. Attribute names and values are extracted to build candidate sets. Frequency estimation is done for each candidate set values. Five sub modules are designed under the data preprocessing module. Transaction view shows the contents of the transaction table. Data cleaning is applied to remove noise records.Attribute list form displays the attribute names and their details. Candidate set preparation module is used to prepare candidate set using attribute names and values. Frequency assignment module is used to estimate frequency values for candidate sets. Frequency values are maintained in separate table.

ONTOLOGY ANALYSIS

We use the ontology is a repository used to maintain the relationship between the concepts and terms. The ontology are maintained as XML documents. The resource description framework is used to manage ontology values. Three types of relationships are represented under ontology. Synonym, meronym and hypernym relationships are used in the ontology. The transaction table attribute names are analyzed with ontology elements. The relationship and their levels are extracted from the ontology analysis. Ontology view shows the concepts with term relationship. Attribute name relationship is produced under attribute analysis form.

RULE MINING PROCESS

The association rule mining tasks are carried out under the rule mining process. Candidate generation is performed with attribute names and attribute values for each transaction. The item sets are prepared from the candidate set information. Frequency values are estimated for each items. The support and confidence values are estimated for all items. The interested rule selection process is carried out on the estimated support and confidence values. Minimum support and minimum confidence values are used to filter the relevant rules.

WEIGHTED RULE MINING

The association rule mining process uses the frequency values for the mining process. It is not suitable for all types of transactions. Weight is used in the rule mining process. Weight and frequency values are used to estimate weighted support and weighted confidence values. The minimum support and minimum confidence values are used to filter the weighted rules.

RULE SELECTION PROCESS

The rule selection process is done with ontology analysis and pruning model. The user assisted rule selection is also carried out to filter the rules. The system selects the rules under the post mining process. The ontology is used extract relationship between the attributes. The rules are ranked with reference to the concept weight values.

CONCLUSION

The proposed system is designed to perform post mining on derived rules. ARIPSO scheme is enhanced with validation methods. Weighted rule mining and filtering process can be integrated with the ARIPSO scheme. Rank based concept relationship analysis can be provided to improve the post mining process. The system is designed to perform rule mining and rule selection process. Ontology is used to reduce the rules based on concept relationships. Weighted rule mining scheme is also integrated with the system.

REFERENCES

[1] Claudia Marinica and Fabrice Guillet"Knowledge-Based Interactive Postminingof Association Rules Using Ontologies" IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 6, June 2010.

[2] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "Mafia: A Maximal Frequent Itemset Algorithm," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 11, pp. 1490-1504, Nov. 2005.

[3] J. Li, "On Optimal Rule Discovery," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 460-471, Apr. 2006.

[4] M.J. Zaki, "Generating Non-Redundant Association Rules," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 34-43, 2000.

[5] B. Baesens, S. Viaene, and J. Vanthienen, "Post-Processing of Association Rules," Proc. Workshop Post-Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics with Sixth ACM SIGKDD, pp. 20-23, 2000.

[6] J. Blanchard, F. Guillet, and H. Briand, "A User-Driven and Quality-Oriented Visualization for Mining Association Rules," Proc. Third IEEE Int'l Conf. Data Mining, pp. 493-496, 2003.

[7] M. Zaki, "Mining Non-Redundant Association Rules," Data Mining and Knowledge Discovery, vol. 9, pp. 223-248, 2004.

[8] M. Hahsler, C. Buchta, and K. Hornik, "Selective Association Rule Generation," Computational Statistic, vol. 23, no. 2, pp. 303-315, Kluwer Academic Publishers, 2008.

[9] A. Berrado and G.C. Runger, "Using Metarules to Organize and Group Discovered Association Rules," Data Mining and Knowledge Discovery, vol. 14, no. 3, pp. 409-431, 2007.

[10] M.J. Zaki and C.J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemset Mining," Proc. Second SIAM Int'l Conf. Data Mining, pp. 34-43, 2002.

[11] A. An, S. Khan, and X. Huang, "Objective and Subjective Algorithms for Grouping Association Rules," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), pp. 477-480, 2003.