International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:1, 2009

Incremental Mining of Shocking Association Patterns

Eiad Yafi, Ahmed Sultan Al-Hegami, M. A. Alam, and Ranjit Biswas

***Abstract*—**Association rules are an important problem in data mining. Massively increasing volume of data in real life databases has motivated researchers to design novel and incremental algorithms for association rules mining. In this paper, we propose an incremental association rules mining algorithm that integrates shocking interestingness criterion during the process of building the model. A new interesting measure called shocking measure is introduced. One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented. The incremental model that reflects the changing data and the user beliefs is attractive in order to make the over all KDD process more effective and efficient. We implemented the proposed approach and experiment it with some public datasets and found the results quite promising.

***Keywords*—**Knowledge discovery in databases (KDD), Data mining, Incremental Association rules, Domain knowledge, Interestingness, Shocking rules (SHR).

I. Introduction

ASSOCIATION rule mining is one of the most important techniques of data mining. It was first introduced in [11]. It aims to extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases. The task of association rules mining usually performed in a two step process. The first step aims at finding all *frequent* itemsets that satisfy the minimum support constraint. The second step involves generating association rules that satisfy the minimum confidence constraint from the frequent itemsets. Since finding the frequent itemsets is of great computational complexity, the problem of mining association rules can be reduced to the problem of finding frequent itemsets.

One of the main drawbacks with the classical association rules algorithms is that they do not consider the time in which the data arrive. In practice, data is acquired in small batches over the time. In such scenario a combination of old and new data is used to build a new model from scratch.

As time advances, some old transactions may become obsolete and thus are discarded from the database. Consequently, some previously discovered knowledge (PDK) becomes invalid while some other new rules may show up. Researchers therefore have been strongly motivated to propose techniques that update the association rule model as new data arrives, rather than running the algorithms from scratch [1,2,3], resulting in incremental models.

Incremental algorithms build and refine the model as new data arrive at different points in time, in contrast to the traditional algorithms where they perform model building in batch manner [1,3]. The incremental association rules algorithms that reflect the changing data trends and the user beliefs are attractive in order to make the over all KDD process more effective and efficient.

We propose an incremental algorithm based on the premise that unless the underlying data generation process has changed dramatically, it is expected that the rules discovered from one set are likely to be similar (in varying degrees) to those discovered from another set [2]. Interesting measures can be used as an effective way to filter the rule set discovered from the target data set thereby, reducing the volume of the output. Our work extends the approaches presented in [4,19] and integrates it into an association rule algorithm in an incremental manner. The proposed approach is a self-upgrading filter that keeps *known knowledge* (previously discovered knowledge (*PDK*) and the user domain knowledge (*DK*) rule base updated as new shocking rules discovered. The shocking interestingness presented in [19] is quantified on the basis of determining significant attributes and then the degree of shocking rules (SHR) of the newly discovered rules with respect to the known knowledge. The idea of shocking rules came from the latest disasters which have encountered the world recently, such as the increasing number of earthquakes, tornados and Tsunami waves. The more interesting rules are those which are unexpected and novel as well, so shocking rules have the highest degree of interestingness. They are novel since they do not exist in the previously discovered knowledge (PDK), unexpected as they have the highest degree of significant attributes that indicates shocking rules (SHR) with respect to the rules in PDK and actionable as they enable the decision maker to make actions to their advantages. A rule is shocking if it overthrows all the expectations of the user. It's unprecedented, never expected and happens suddenly in a way that it shocks the user and put him in an unenviable situation.

The proposed incremental association rule algorithm operates on the incremental training set and builds a model. During the frequent itemsets generation, the algorithm computes the shocking interesting measure against the *known knowledge* and prunes the items that do not meet the user

Eiad Yafi is a research scholar from Syria, working on his PhD in Data Mining at Hamdard University, New Delhi (e-mail: eiad.yafi@ gmail.com).

Ahmed Sultan Al-Hegami is assistant professor of Artificial Intelligence and Intelligent Information Systems, Sana'a University, Yemen (e-mail: aal-hegami@suye.ac).

M.A.Alam is a professor at Hamdard University, New Delhi (e-mail: alam@jamiahamdard.ac.in).

Ranjit Biswas is a visiting professor at Hamdard University, New Delhi (e-mail: ranjitbiswas@yahoo.com).

interest. A detailed description of computation of shocking interestingness measure can be found in [19]. Further iteration of the algorithm is performed only for the frequent itemsets that have shocking interestingness measure higher than a user specified threshold. This is a useful feature in which the user may need to trade off some accuracy for shocking interestingness that may arise in some domains where the user wants a rough picture about the domain rather than an optimal model that contains a lot of details.

The incremental nature of the proposed algorithm makes it advantageous to discover shocking patterns at current time with respect to the previously discovered patterns (rules), rather than exhaustively discovering all patterns.

II. Related Work

Several approaches have been proposed for developing incremental algorithms of association rules mining [13,14,15,16,17,18] and mining of the frequent itemsets [16,17,18]. The main assumption of these approaches is to update the discovered model when new data arrive. In [13], DEMON algorithm is proposed that works effectively and efficiently with evolving data over the time. [14] proposed algorithm for monitoring the changes in the data stream environments [9]. Another incremental algorithm, Lee and Cheung uses statistical methods for updating process with DELI algorithm. DELI algorithm applies a sampling technique to estimate the support counts using an approximate upper/lower bounds on the amount of changes in the set of newly introduced association rules. A low bound would mean that changes in association rules are small and there should be no maintenance. In addition, these algorithms are incremental in nature as they use and reuse the previously discovered knowledge and integrated when new data occur. Zaki et al designed a parallel approach that reduces the computational requirement in the algorithm [15].

Cheug et al proposed FUP (Fast Update) algorithm for the maintenance of discovered association rules in large databases. They have proposed to handle incremental database by scanning the database to check whether there are large itemsets or not. FUP algorithm is introduced for quantifying the large itemsets in the updated database. The goal of this algorithm is to solve the efficient update problem of association rule in updated database. They have extended their work to FUP* and FUP2 to k-pass algorithms, they scans the database k^{th} time [17,18].

In this paper, we propose a new approach to dynamic mining association rules in evolving databases. The proposed approach integrates the shocking interestingness criterion [19] during the model building process to form a constraint in order only to discover shocking rules.

The rest of the paper is organized as follows. The problem statements are given in section III. Section IV presents the shocking interestingness measure used in the proposed algorithm. In section V, our proposed approach for incremental generation of association rules is introduced. Section VI presents the algorithm used in the proposed approach. Section VII shows the effectiveness of the proposed approach through a detailed example. Experimental results and implementation are presented in section VIII. Finally, section IX contains the conclusion.

III. Problem Statement

Given a dataset D collected over the time $[t_0,t_1,t_2,\ldots t_n]$. At t_0, D_0 represents an empty database. At time instance t_i, an incremental dataset D_i , $i \in \{1,\ldots,n\}$, is collected such that $D = D_1 \cup D_2 \cup \ldots D_i$. Let T_i and T_{i+1} be two models discovered at time instances t_i and t_{i+1} from datasets $\bigcup_{j=1}^{i} D_j$ and $\bigcup_{j=1}^{i+1} D_j$ respectively. The major volume of discovered rules in T_i and T_{i+1} would be similar to some extent. A small set of rules, which are either present or absent in T_{i+1} represents change in data characteristics. The objective is to update T_i to T_{i+1} using D_{i+1} and T_i. T_i – the model discovered at time t_i now represents PDK. T_{i+1} is the up-to-date model obtained by adding interesting rules discovered from D_{i+1}. This is achieved by constructing a model $\check{T}_{i+1}$ from D_{i+1} such that association rules in $\check{T}_{i+1}$ have user specified degree of shocking interestingness with respect to the rules in T_i. Subsequently, $\check{T}_{i+1}$ is used to update T_i to T_{i+1}.

IV. Shocking Interestingness Measure

The shocking interestingness of a rule is quantified by computing the significant attributes and then the degree of shocking rules (SHR) of the antecedent and the consequent at conjunct level and subsequently the significance at conjunct level is combined to compute the significance at rule level. A high significant attributes indicates a high degree of shocking rules (SHR). The significant attribute is computed by measuring the deviation for the antecedent and the consequent at conjunct level and subsequently, combine the conjunct level deviation to compute rule level deviation. The detailed description of our proposed approach to quantify shocking interestingness can be found in [19].

In this work we integrate the approach proposed in [19] into the mining association rule algorithm. The shocking measure is pushed into the classical Apriori algorithm to form a constraint in order to discover only shocking interesting patterns. The shocking measure is applied at each iteration, of the frequent itemsets generation. For each itemset, a set of rules are extracted to form strong partial rules. The strong partial rules are those rules with min_confidence. These partial rules are subjected to the shocking interestingness criterion resulting in computation of the degree of shocking interestingness which is assigned to each partial rule. The partial rules are shocking if the degree of their interestingness is higher than a user interestingness threshold value.

The process of frequent itemsets generation is further continuing taking into account the rules that are only interesting and pruning the itemsets that are uninteresting. This strategy ensures that only shocking interesting itemsets are eligible to be candidate during the next iteration of frequent itemsets generation.

V. Incremental Mining of Association rules

The proposed algorithm uses shocking interestingness measure as a constraint during the model building in order to discover association rules that are shocking (interesting) for the user. The advantage of pushing such constraints inside the algorithm is that the search space is reduced and the algorithm discovers relatively smaller sized model. Further, the discovered knowledge reflects the user's requirement of interestingness.

One of the main features of the proposed algorithm is to deal with time changing data and user beliefs. This is a useful functionality in situations when two datasets have arrived at different points of time or from different geographical locations. Certainly, it is attractive to update the discovered knowledge each time new data arrive.

For every stage of frequent itemsets generation, partial rules[1] are generated from the frequent itemsets at that stage. These partial rules are evaluated using confidence measure and prune the partial rules that do not satisfy this measure resulting in a set of strong partial association rules. The strong rules are subjected to the shocking criterion [19] in order to decide either these rules are interesting or not. The shocking interestingness (SI) of a partial rule *P* is computed with respect to the closest rule *R* in the existing model T_i. The $SI_P^R = 0$, indicates that the partial rule if expanded is likely to create a rule which is already present in T_i. If the partial rule is found to be shocking interesting i.e. $SI_P^R = 1$, the algorithm expands the current frequent itemsets to next level frequent itemsets like a normal *Apriori* algorithm. In case a partial rule is found to be uninteresting, it computes the SIGNIFICANCE *factor* (SF) of the partial rule P to decide whether the partial rule is to be expanded further or not. Computation of SIGNIFICANCE *factor* of P indicates the relevance of rule R with respect to current training set (D_{i+1}). If the SIGNIFICANCE *factor* is acceptable, the partial rule is not expanded further.

Fig. 1 shows the environment in which the proposed algorithm operates. At time t_{i+1}, database D_{i+1} is pre-processed and subjected to the algorithm. The algorithm takes into account the existing model T_i representing the known association rules. The algorithm expands only those frequent itemsets that are likely to lead to shocking interesting rules. This results into discovering of $\check{T}_{i+1}$. For each frequent itemsets, a rule is extracted and used to update the model T_{i+1}.

[1] We use the term partial rule because the complete association rule has not been generated yet, and addition of items is certain to be added to the rules.

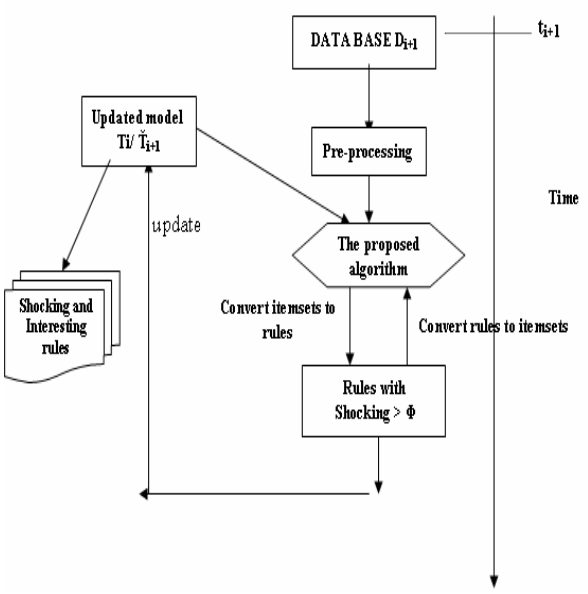

Fig. 1 Operating environment of the proposed algorithm

A. Frequent Itemsets Generation

The proposed algorithm is based on the generic Apriori methodology [11]. It dynamically decides whether an itemset is to be used in the next iteration of candidate generation or not taking into account the partial rule that is obtained by converting the frequent itemsets, which have confidence higher than confidence threshold value, into rules.

At each frequent itemsets, the following tasks are performed.

1. Extracting the partial rules which have confidence higher than the confidence threshold value,
2. Computation of shocking interestingness (SI) of the partial rules P with respect to the existing model T_i,
3. Computation of SIGNIFICANCE factor of the partial rule P with $SF_P^R = 0$.

The process of candidate generation is continued as the tradition Apriori algorithm taking into account the only shocking interesting frequent items.

B. Dynamic Pruning Based on Shocking Interestingness

The characteristic feature of the approach is its ability to facilitate dynamic pruning based on shocking interestingness [19]. The objective is to reduce the size (complexity) of the frequent itemsets generation with assurance that the resulting rules does not compromise in terms of *accuracy* and provides the user with shocking interesting association rules**.**

The algorithm computes shocking interestingness (SI) at each iteration of frequent itemsets generation to determine whether an itemset is likely to lead to an interesting rule or not. An itemset becomes a candidate for next level frequent itemsets generation if its shocking interestingness (SI) value is one or the significance factor of the closest rule in T_i is less than the significance factor threshold value. An interestingness value of 1, of the partial rule indicates that this rule is unlikely to expand to any existing association rule. A shocking interestingness value (SI) of 0 of the partial rule indicates that the partial rule is likely to expand to some

existing association rule. The threshold used to compute shocking interestingness can be set dynamically according to the user requirement. This flexibility to dynamically change the threshold is a useful feature in a situation where the user has a rough picture about the domain and is in learning phase.

C. SIGNIFICANCE Factor of a Partial Rule

The algorithm computes the significance factor (SF) at each partial rule P with $SI_P^R = 0$, to judge significance of the rule in the current training set D_i. The computation of significance factor (SF) is required in order to be assured that the expected expansion of the current partial rule is significant in the current increment D_{i+1} of the database. A higher significance factor of the expected expansion of a partial rule with respect to the training data indicates that the rule is still significance at current time t_{i+1}. A smaller significance factor, on the other hand, indicates that the expected expansion of the partial rule is now obsolete and is not valid in this dataset. As a result, the itemset must be further expanded and may lead to an interesting rule.

The significance factor computation is done using the following methodology.

Given a rule $A \rightarrow R$, the subset of the training set corresponding to A is called cover of A (Γ_A). The significance factor (SF) of $A \rightarrow R$ is given as follows:

$$SF\ (A \rightarrow R) = \frac{|\Gamma(A \cap R)|}{|\Gamma_A|}$$

where $|\Gamma(A \cap R)|$ denotes the number of tuples that contain both A and the class R, and $|\Gamma_A|$ is the number of tuples that contain antecedent A.

Let R^P be the partial rule and $R^S(A \rightarrow R)$ be the closest rule in T_i such that $SI(R^P, R^S) = 0$. Then

$$SF(R^P) = \frac{|\Gamma(A \cap R)|}{|\Gamma_A|}$$

Having computed the significance factor of the partial rule, the algorithm expands the itemset if the significance factor is lower than the specified significance factor threshold value and stops expanding otherwise.

VI. Experimental Study

The proposed approach is implemented and tested using several public datasets available at http://kdd.ics.uci.edu. The approach is implemented using c programming language. The datasets are partitioned into three groups representing instances arrive at time T_1, time T_2 and T_3 respectively with 0.1% and 1% to indicate minimum confidence and minimum support respectively. The proposed approach uses the discovered rules to find out those rules which are interesting and which are conforming. The conforming rules are the rules that confirm the user background knowledge and hence are not interesting.

A. Experiment

The objective of the first experiment is to show the effectiveness of our approach in reducing the number of discovered rules. It is expected that the number of discovered rules keeps on decreasing over the time. We work with five datasets and assume that the shocking interestingness threshold value (SI) = 0.6. The values in the third column of Table 2 represent the number of rules discovered, using WEKA, at a given partition and the values in the fourth column represent the shocking interesting rules discovered by our approach. It is observed that the number of interesting rules decreases in contrast to the number of conforming rules which increases each time new rules discovered as per our expectation. Intuitively, the interesting rules discovered at time T_1 is no more interesting at time T_2. The conforming rules are at the last column of Table II.

TABLE II
THE DISCOVERED MEDICAL RULES AT TIME T_1, T_2, AND T_3

Dataset	Time	Discovered AR's	Interesting rules	Conforming rules
Lymph	T_1	32000	18230	13770
	T_2	28562	12003	16559
	T_3	26781	2010	24771
Breast	T_1	802	320	482
	T_2	725	180	545
	T_3	540	73	467
Hepatitis	T_1	1207	800	407
	T_2	980	430	550
	T_3	626	228	398
Heart	T_1	987	564	423
	T_2	566	320	246
	T_3	207	118	89
Sick	T_1	4502	2876	1635
	T_2	2709	1078	1631
	T_3	986	401	585

VII. Conclusion

In this paper, we have proposed an algorithm for incremental association rules mining that integrates shocking interestingness criterion during the process of building the model. A new interesting measure called shocking measure is introduced. One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented The proposed approach is a self-upgrading filter that utilizes interestingness criterion to reflect the user subjectivity and extract patterns, incrementally, from datasets arrive at different points in time. The proposed algorithm makes use of interestingness measure as the basis of extracting interesting patterns. This important feature of the proposed algorithm is attractive and desirable in many real life applications as the volume of data keeps on growing and changing over the time and therefore the user background knowledge is monotonically augmented. This changing environment updates the user understandability and comprehensibility about the domain.

REFERENCES

[1] Han, J. and Kamber, M.: Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers, (2001).
[2] Dunham M. H.: Data Mining: Introductory and Advanced Topics. 1st Edition Pearson Education (Singapore) Pte. Ltd. (2003).
[3] Hand, D., Mannila, H. and Smyth, P.: Principles of Data Mining, Prentice-Hall of India Private Limited, India, (2001).
[4] Kaur H., Wasan. S. K, Al-Hegami A. S., Bhatnagar, V.: A Unified Approach for Discovery of Interesting Association Rules. To appear in Proceedings of Industrial Conference on Data Mining (ICDM), 2006.
[5] Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Adaptive Constraint Pushing in Frequent Pattern Mining. In Proceedings of the 17th European Conference on PAKDD03. (2003).
[6] Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: ExAMiner: Optimized Level-wise Frequent pattern Mining with Monotone Constraints. In Proceedings of the 3rd International Conference on Data Mining (ICDM03). (2003).
[7] Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Exante: Anticipated Data Reduction in Constrained Pattern Mining. In Proceedings of the 7th PAKDD03. (2003).
[8] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A. I.: Finding Interesting Rules from Large Sets of Discovered Association Rules. In Proceedings of the 3rd International Conference on Information and Knowledge Management. Gaithersburg, Maryland. (1994).
[9] Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems. (2000).
[10] Psaila, G.: Discovery of Association Rules Meta-Patterns. In Proceedings of 2nd International Conference on Data Warehousing and Knowledge Discovery (DAWAK99). (1999).
[11] Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases, In ACM SIGMOD Conference of Management of Data. Washington D.C., (1993).
[12] Hansel, G.: Sur le nombre des functions Boolenes Monotones den variables. C.R. Acad. Sci. Paris, 262(20):1088-1090 (in French). (1966).
[13] Ganti, V., Gehrke, J. and Ramakrishnan, R.: DEMON: Mining and Monitoring evolving data. In Proceeding of the 16th International Conference on Data Engineering, San Diego, USA. (2000).
[14] Lee, S., and Cheung, D.: Maintenance of discovered association rules. When to update? In Research Issues on Data Mining and Knowledge Discovery. (1997).
[15] Zaki, M. and Hsiao, C.: Charm: An efficient algorithm for closed itemset mining. In Proceeding of the 2nd SIAM International Conference on Data Mining, Arlington, USA. (2002).
[16] Cheung, D. W., Han, J., Ng, V.T., Wong, C.Y.: Maintenance of discovered Association Rules in Large Databases: An Incremental Updating Technique, Proc. the International Conference On Data Engineering, (1996) 106-114.
[17] Cheung, D. W., Ng, V.T., Tam, B.W.: Maintenance of Discovered Knowledge: A case in Multi-level Association Rules, Proc. 2nd International Conference on Knowledge Discovery and Data Mining, (1996) 307-310.
[18] Cheung, D. W., Lee, S.D., Kao, B.: A general Incremental Technique for Mining Discovered Association Rules, Proc. International Conference on Database System for Advanced Applications, (1997) 185-194.
[19] Yafi, E., Alam, M.A., Biswas, R.: Development of Subjective Measures of Interestingness: From Unexpectedness to Shocking, Proceedings of World Academy of Science, Engineering and Technology Volume 26 December 2007 ISSN 1307-6884.