IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 1, Feb-Mar, 2014
ISSN: 2320 - 8791
www.ijreat.org

Discrimination In Combined Mining For Informative Knowledge And Automated Data

V.Kaleeswari[1], N.Abirami[2]
[1]**Computer Science and Engineering, Sree Sastha Institute of Engineering and Technology, Chennai, Tamilnadu, India**
[2]**Computer Science and Engineering, Sree Sastha Institute of Engineering and Technology, Chennai, Tamilnadu, India**

Abstract

Data mining application involve multiple large data sources and business application and user preferred manually. A single method or one-step mining is often limited in discovering informative knowledge. It would also be very time and space consuming, if not impossible, to join relevant large data sources for mining patterns consisting of multiple aspects of information. Automated data collection and data mining techniques such as classification rule mining have covered the way to making automated decisions. An effective approaches for mining patterns combining necessary information from multiple relevant business lines. In this way discrimination occur when decisions are made sensitive attributes. So we propose combine mining approach to combined informative patterns and combined multiple data sets or by multiple methods. They identify combined pattern and improving services and which show the flexibility of combined informative knowledge in automated data for efficient way of approach.

Keywords: Data mining combined mining, data sources, automated data, knowledge discovery data, association rules.

1. Introduction

Data mining applications, such as mining public service data, inevitably involve complex data sources, particularly multiple large scale, distributed, and heterogeneous data sources embedding information about business transactions, user preferences, and business impact. In these situations, most of business oriented people certainly expect the discovered knowledge to present a full picture of business settings rather than one view based on a single source. It is challenging to mine for comprehensive and informative knowledge in such large automated data suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining.

In developing effective techniques for involving multiple heterogeneous features, data sets, and methods in enterprise data mining and Services in the information society allow for automatic and routine collection of large amounts of data. One must prevent data mining from becoming itself a source of discrimination, due to data mining tasks generating discriminatory models from biased data sets as part of the automated decision making. The existing works in handling the challenges can be categorized into the following aspects: 1) data sampling, 2) involving multiple methods; and 3) mining multiple data sources. In these techniques for involving multiple methods and handling multiple data sources are often specifically developed for particular cases.

The concepts of combined association rules, combined rule pairs, and combined rule clusters to mine for informative patterns in complex data by catering for the comprehensive aspects in multiple data sets. A combined association rule is composed of multiple heterogeneous itemsets from different data sets while combined rule pairs and combined rule clusters are built from combined association rules. Analysis shows that such combined rules cannot be directly produced by traditional algorithms such as the FPGrowth. The existing works and proposes the approach of combined mining as a general method for directly identifying patterns enclosing constituents from multiple sources or with heterogeneous features such as covering demographics, behavior, and business impacts. Its deliverables are combined patterns such as the aforementioned combined association rules. Combined patterns consist of multiple components, a pair or cluster of atomic patterns, identified in individual sources or based on individual methods.

The general ideas of combined mining are as follows:
1) By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.

2) By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.

3) By applying multiple methods in pattern mining, combined patterns are generated which disclose a deep

and comprehensive essence of data by taking advantage of different methods.

4) By applying multiple interestingness metrics in pattern mining, patterns are generated which reflect concerns and significance from multiple perspectives.

2. Related work

First, most of existing single-handed data mining methods do not target the discovery of informative patterns in automated data For instance, a combined association rule R is in the form of R: A1 . . . ⋀ Ai ⋀ B1 ⋀ . . . Bj → T, where Ai ∈ Di and Bj ∈ Dj are itemsets in heterogeneous data sets Di and Dj , respectively, T is a target item or class. Analysis shows that such combined rules cannot be directly produced by traditional algorithms such as the FPGrowth.

Second, approaches to mining for more informative and actionable knowledge in large data can be generally categorized as follows: 1) direct mining by inventing effective approaches; 2) involving extra features from other data sets; 3) integrating multiple methods; and 4) joining multiple relational tables. Direct mining for discriminative patterns has been highlighted data set is a collection of data objects (records) and their attributes. Let DB be the original data set. An item is an attribute along with its value, e.g., Race =black. The integration of multiple data mining methods is widely used to mine for more informative knowledge, such as associative classification, combining clustering and association rules for rarity mining combining regression with association rule mining, and association-rule-mining-based classification.

Table 1: Customer details

CustomID	*Gender*
1	*F*
4	*M*
3	*F*

Table 2: Traditional Association Rules

Rules	*Support*	*Confidence*	*Lift*
C1→Y	*6/10*	*4/6*	*1.4*
C2→N	*5/10*	*6/8*	*0.7*
C1→Y	*4/10*	*5/6*	*1*
C2→N	*8/10*	*3/8*	*1*

Table 3: Combined Association Rules

Rules	*Support*	*Confi*	*Lift*	*Count*
F ∧ c1 →Y	*6/10*	*1/2*	*1.3*	*1*
M ∧ c1→N	*5/10*	*3/4*	*1.4*	*1.2*
F ∧ c2 →Y	*4/10*	*4/5*	*1*	*1.3*
M ∧ c2→N	*2/10*	*3/5*	*0.3*	*0.6*

Table joining is widely used in order to mine patterns from multiple relational tables by putting relevant features from individual tables into a consolidated one. As a result, a pattern may consist of features from multiple tables. This method is suitable for mining multiple relational databases, particularly for small data sets. Combined mining can identify such compound patterns in large data sets. Multi relational data mining and multi database mining have been intensively studied. They are different from combined mining. The multisource combined mining shows; combined mining does not rely on joining related tables. The resulting patterns of the multisource combined mining can consist of pairs or clusters of patterns with components from multiple data sets, which is new to multi relational mining, to the best of our knowledge.

2.1 Combined Association Rules

The task involves a large scale of real-world complex public service data, including customer information in Table I, unordered government policies applied on customers, ordered customer activities, and the impact of customers on government service objectives, namely, whether a customer incurs debts or not. For instance, when association mining is used to mine frequent rules, the rules shown in Table II can be discovered from the unordered transactional data set.

In combined mining to produce more informative and actionable patterns. These are:

1. The whole population into male and female groups, based on the demographic data in Table I, and then mines the demographic and transactional data of the two groups separately, as partially shown in Table III, where Cont denotes the contribution of the transactional data and I rule reflects the interestingness of the combined rules.

2. Frequent patterns combining unordered and ordered items can be identified and from Table III the male customers under policy $c2$ are very likely to have a debt since its I pair is as high.

2.2 Combined mining pattern

Combined mining is a two-to-multistep data mining procedure and it involves types, structures formed by atomic patterns, and relationships and time frames among atomic patterns.

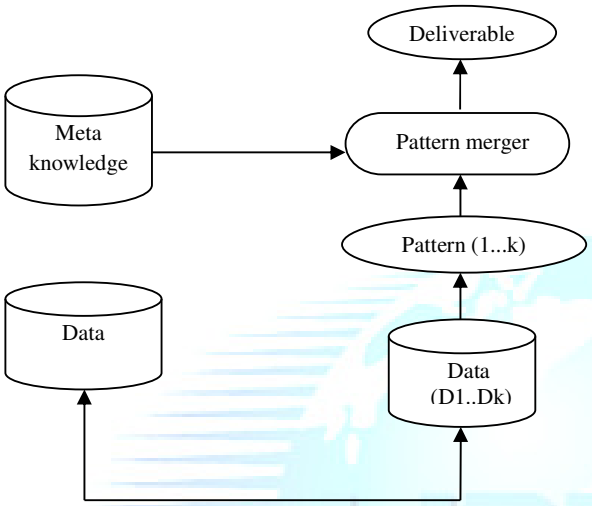

Fig .1. Basic functions in combined mining

Fig. 1 illustrates a framework for combined mining. It supports the discovery of combined patterns either in multiple data sets or subsets (D1, D2…, and DK) through data partitioning in the following manner:

1) Based on domain knowledge, business understanding, and goal definition, one of the data sets or certain partial data (say *D*1) are selected for mining exploration,

2) The findings are used to guide either data partition or data set management through the data coordinator and to design strategies for managing and conducting serial or parallel pattern mining on relevant data sets or subsets or mining respective patterns on relevant remaining data sets. The understanding of the data/business and objectives, and if necessary, another step of pattern mining is conducted on data set Dk with the supervision of the results from step k – 1; and

3) After finishing the mining of all data sets, patterns identified from individual data sets are merged with the involvement of domain knowledge and further extracted into final deliverables (P).

3. The Algorithm

Discrimination prevention method algorithm involved in the automated data sets and discrimination item sets and Method 1 for DRP. For each direct discriminatory rule r0 in MR (Step 3), after finding the subset DBc (Step 5), records in DBc should be changed until the direct rule protection requirement (Step 10) is met for each respective rule (Steps 10-14).

Algorithm

DIRECT RULE PROTECTION METHOD

1: Inputs: DB, FR, MR, _, DIs
2: Output: DB' (transformed data set)
3: for each r': A, B → C ϵ MR do
4: FR →FR - f{r'}
5: DBc ←All records completely supporting ┐ A, B→┐ C
6: for each dbc ϵ DBc do
7: Compute impact (DBc→FR) supports the premise of RJ
8: end for
9: Sort DBc by ascending impact
10: while conf (r') ≥ conf (B → C) do
11: Select first record in DBc
12: Modify discriminatory item set of dbc from ┐ A to A in DB
13: Recompute conf (r')
14: end while
15: end for
16: Output: DB'= DB

Among the records of DBc, one should change those with lowest impact on the other (protective or none redlining) rules. Hence, for each record DBc, the number of rules whose premise is supported by dbc is taken as the impact of dbc (Step 7), that is impact dbc'; the rationale is that changing dbc impacts on the confidence of those rules. Then, the records dbc with minimum impact dbc' are selected for change (Step 9), with the aim of scoring well in terms of the utility measures proposed in the next section. Then call this procedure (Steps 6-9) impact minimization. Rule generalization should be performed (Step 5), after determining the records that should be changed for impact minimization (Steps 7-8), these records should be changed until the rule generalization requirement is met (Steps 9-13). Also if shows that direct rule protection should be performed (Step 15), based on either Method 1 in the automated data.

4. Evaluation of methods

4.1 Mining Multi method-Based Combined Discrimination Method Patterns

In conduct sequence classification based on identified frequent positive and negative sequences. These analyze the relationship between the transactional activity patterns and the debt occurrences and build sequence classifiers for debt detection. Negative sequential rules have been used to find both the positive

and negative sequences in the Centre link customer debt-related activity data. To measure data quality, and use two metrics proposed in the literature as information loss measures in the context of rule hiding for privacy-preserving data mining (PPDM).

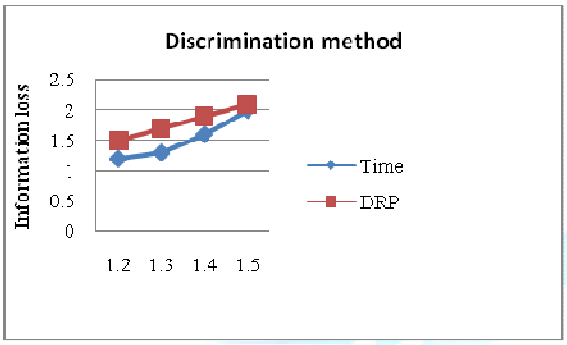

Fig.2 Information loss in discrimination method

Fig.2 is the effect of the discrimination method, described in on execution times and information loss of Method 1 for DRP, respectively. As shown in this figure (right) impact minimization has a noticeable effect on information loss to show the effect of varying the minimum support and the minimum confidence on the proposed techniques.

5. Conclusion and Future Work

Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than obvious that most people do not want to be discriminated. Building on existing works, this paper has presented a comprehensive and general approach named combined mining for discovering informative knowledge in complex data. It will focus on discussing the frameworks for handling multi feature-, multisource-, and multi method-related issues. They have shown that the proposed frameworks are flexible and customizable for handling a large amount of complex data involving multiple features, sources, and methods as needed, for which data sampling and table joining may not be acceptable. They have also shown that the identified combined patterns are more informative and actionable than any single patterns identified in the traditional way.

Further developing effective paradigms, combined pattern types, combined mining methods, pattern merging methods, and interestingness measures for handling large and multiple sources of data available in our industry projects and the proposed techniques are quite successful in both goals of removing discrimination and preserving data quality. The perception of discrimination, just like the perception of privacy, strongly depends on the legal and cultural conventions of a society.

6. References

[1] Y. Zhao, C. Zhang, and L. Cao, Eds., Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction. Hershey, PA: Inf. Sci.Ref., 2009.

[2] A. Jorge "Hierarchical clustering for thematic browsing and summarization of large sets of association rules," in Proc. SDM, 2004, pp. 178–187.

[3] M. Plasse, N. Niang, G. Saporta, A. Villeminot, and L. Leblond, "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set," Comput. Statist. Data Anal., vol. 52, no. 1, pp. 596–613, Sep. 2007.

[4] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Combined association rule mining," in Proc . PAKDD, 2008, pp. 1069–1074.

[5] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination- Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[6] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.

[7] Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang, "Mining for combined association rules on multiple datasets," in Proc. DDDM, 2007, pp. 18–23.

[8] Y. Zhao, H. Zhang, S. Wu, J. Pei, L. Cao, C Zhang, and H. Bohlscheid, "Debt detection in social security by sequence classification using both positive and negative patterns," in Proc. ECML-PKDD, vol. 5782, LNAI, 2009, pp. 648– 663.