Combined Mining process on complex data

Mayuri Mohan Tharval
Information Technology Department
Padmabhushan Vasantdada Patil Pratisthan College of Engineering
Mumbai, India

Abstract – The data mining and knowledge discovery in databases have been the latest trend for researchers, media attention, and industry. The stake of the people has grown tremendously in business to analyze the data in all sectors. It has discovered the ways of using the data scattered as one system for better understanding and management of the business. The Enterprise data mining consist of applications which includes complex data such as multiple heterogeneous large data sources, business impact and user preferences. For this situation single method is often limited in discovering informative knowledge from complex database. It is critical to develop effective methods for mining patterns combining necessary information from multiple related business rules, catering for real business needs and decision-making actions rather than just providing a single line of patterns. The mechanism describes the basic processes for complex data like multi method combined mining, multi feature combined mining and multi source combined mining. The mechanism will be evaluated with complex data sets to show the flexibility and instantiation capability in discovering informative knowledge in complex data using combined mining.

Keywords - Data Mining, knowledge discovery, complex data, combined mining, multimethod, multisource, multifeature.

I. INTRODUCTION

The data mining applications, like mining public service data and telecom fraudulent activities, consist of complex data sources, particularly distributed ,multiple large scale, and heterogeneous data sources embedding information about user preferences ,transactions, and business impact. In these applications business people are discovering the informative knowledge to prevent the business settings on a single sources object. At this situation business people think the need to discovering the informative knowledge by using multiple business rules with effective and high performance. This process is a challenge and very difficult by applying the business rules in the joined tables and decision makers to get the informative knowledge from complex data. The following few drawbacks of the discovering knowledge information from the complex data:

1) Data sampling,
2) Joining multiple relational tables,
3) Post analysis and mining,
4) Involving multiple methods, and
5) Multiple data sources mining.

In real-life data mining, the data sampling is usually not acceptable as it may miss important data that are filtered out. The joining of tables may not be possible due to the time and space limit such as dealing with hundreds of millions of transactions from multiple data sources. In [5], [4], and [9], we proposed the concepts of combined rule pairs, combined rule clusters, association rule and combined rule pairs. A combined association rule is composed of multiple heterogeneous itemsets from different data sets while combined rule pairs and clusters are built from combined association rules. These combined rules cannot be directly produced by traditional algorithms such as the FPGrowth [13].

The Proposed system in this paper is completely dependent on the existing system because the proposed system uses the existing algorithms and methods are used to develop. We are proposing the combined association rules, combined rule pairs and combined rule cluster are used to mine the informative data from complex data.

II. CONCEPT OF COMBINED MINING

Combined mining is proposed methodology for handling the complexity of employing multi-information sources, multi-feature sets, constraints, multi-methods and multi-models in data mining, and for analyzing the complex relations between descriptors or objects (attributes, sources, methods, constraints, labels and impacts) or between identified patterns during the learning process. Combined patterns can be formed through the analysis of internal relations between pattern constituents or objects that are obtained by method on a single dataset, for which the combined sequential patterns are formed by analyzing the relations within a sequential pattern set. The contribution of combined mining is that it enables the induction of knowledge, construction, extraction and discovery which consists not only of discriminated objects but also of relations between objects and interactions, as well as their impact. This approach is called actionable complex patterns. Combined mining provides a solution for meeting the challenge of mining complex knowledge in complex data [6]. It also builds upon other approaches such as

generalization, summarization [8, 2], and aggregation and inference, in order to combine them with data-driven knowledge discovery from complex environments. It helps to clear the difference and purpose between combined mining and other relevant techniques. First, the actionability of patterns [6, 7] and the method to discover actionable patterns [3, 14, 11] even though we assure that combined mining intends to deliver actionable results. Secondly, combined mining handles a range of different scenarios from multiple sets of features [15] to multiple methods and multiple sources as necessary for problem solving [1]. Finally, combined mining could be treated as a hybrid data mining method if it was to be applied for the multi-method combined pattern mining to execute the variety of combinations of different data mining approaches [12, 10]. The combination of classification with association rule mining, it produces an associative classifier which can be built on an unordered dataset to predict online shopping fraud. Combined mining focuses on extracting and discovering more meaningful patterns from complex data. The flow of the combined mining process is as shown in figure 1.

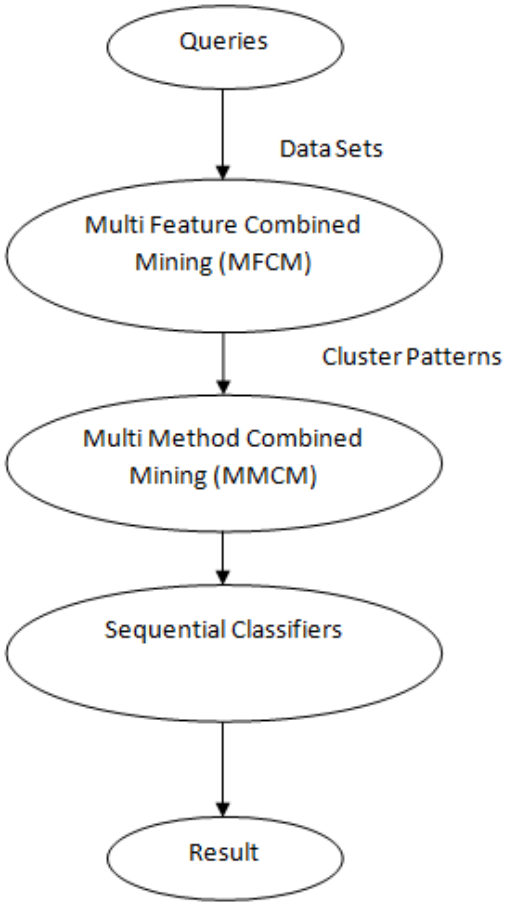

Figure 1. Steps of Combined Mining

III. MULTI SOURCE COMBINED MINING

The various inputs of data mining application come from different data sources. These data sources can be heterogeneous, huge as well as complex in size. Instead of combining this data into homogeneous form (which is usually done by joining the tables) and then operated upon by mining algorithms, it is recommended to combine discovered patterns from the heterogeneous data sources [1].

IV. MULTI FEATURE COMBINED MINING

Multiple data sources with patterns like homogeneous are very easy to handle, but data sources with multiple feature-set required to be handled with special care. In multifeature combined mining, atomic patterns (the patterns that are generated from a single source) are collaborated to form combined patterns which are found to be informative. The steps in MFCM are shown in figure 2.

The data obtained after applying the patterns is informative which helps to make decisions for the growth of the business. The collaboration of the patterns uses business intelligence. But data with multiple features is very delicate and need to be handled with care.

The diagrammatic representation helps to understand properly the concepts of combined mining. The sequential classifiers are used for the final result of combined mining so that further actions can be taken depending on the output. So the input given for mining should be correct and according to the requirements of the user.

The following are some types of combined patterns [1] discussed as:

a) Pair patterns are formed by the combining of two atomic patterns. They are of form, {A1→B1,A2→B2} where A1 and A2 are same but B1 and B2 are not the same or vice versa. The new measure Ipair, is useful for measuring the measure the interestingness of pair pattern.

b) Cluster patterns, are formed by the collection of combining together the similar or related or atomic patterns. They take the form, {A1→B1, A2→B2, A→BN}. Measure Icluster explains how interested the cluster of patterns is.

c) Some patterns can sometimes take form of extension of other patterns. For example, {A1U B1→C1} can be referred as an extension of {A1→C1}. Such patterns are merged to form incremental pair patterns.

d) The patterns which are extension of one another are to be grouped together to form Incremental cluster patterns, as {A1→Z1, A1U B1→Z1, A1U B1U C1→Z1…}.

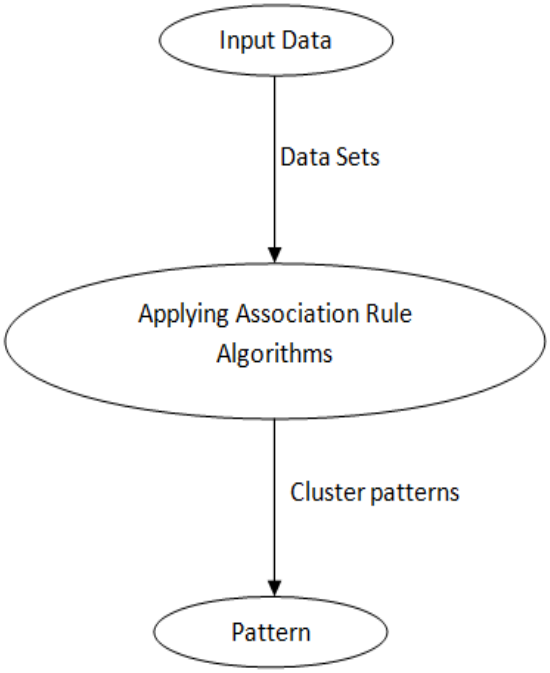

Figure 2. Steps of MFCM

V. MULTI METHOD COMBINED MINING

In many cases the patterns that are discovered by a particular method that do not serve to user's perspective. Here we can find need of using more than one methods of mining in order to discover more informative patterns. The multiple methods can be used serially, closed loop or in parallel fashion [1]. The steps in MMCM are shown in figure 3.

a) Parallel Multimethod Mining: In this type of approach various mining techniques are used parallel on same or different data sets.
1) Firstly the independent data sources are mined by using different data mining techniques available, to find out the respective atomic pattern sets.
2) Then the available patterns are collaborated together by merging method.

b) Serial Multimethod Mining: The various data mining techniques are used here one by one. So the outcome of one technique is handled by another technique to discover the in depth knowledge and this process goes on till the last technique of the process.
This method works as follows:
1) Firstly the data source available is mined to obtain pattern set P1 with the help of suitable method.
2) After studying the initial pattern set, the next suitable method is selected and then P1 is again mined using it to discover the next level pattern P2.
3) The available different techniques are applied according to the domain knowledge and the output requirement.

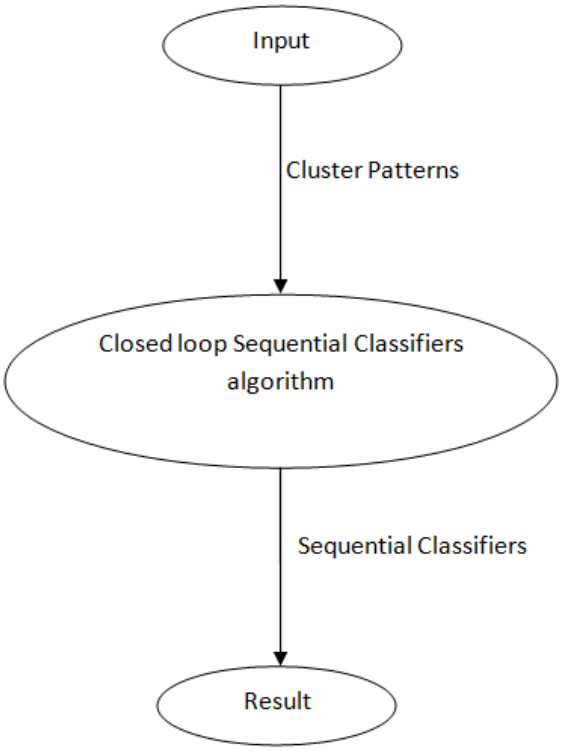

Figure 3. Steps of MMCM

c) Closed Loop Multimethod Mining: This approach considers the impact of one technique on the other. Practically the feedback from next method can be used to purify the results obtained from previous method. So it can be easily observed that in closed loop approach, the decision (whether the pattern is interested or not) depends on other methods also and not only on that particular method. Closed loop mining is consist the following steps which are as follows:
1) Initially the pattern set P1 is formed by following the process of serial multimethod mining.Some samples may not be identified at the end of this step. This is due to the conditions and limitations that are applied on various mining techniques.
2) The patterns in P1 are verified for their validity. Some samples may not be verified to patterns. The separate data set is formed using such uncommon samples, say Dx. This available dataset is again mined by multiple methods to discover the new pattern P2.
3) Process of step 2 is repeated many times as required by the miner to discover patterns like P1, P2 …Pk. This processes help the patterns to rediscover the new data.
4) All patterns are then collaborated to form combined pattern.

VI. CONCLUSION

The typical enterprise applications, like the telecom fraud detection and close observation of stock markets, involve heterogeneous and multiple distributed features. This data sources with huge quantities, expect to cater for preferences, user demographics, business impact, behavior, business appearance and service usage. There is a need to mine for patterns consisting of various aspects of the information which reflects comprehensive business background. This present patterns help in decision-making actions. The challenges of the existing data mining methods like postanalysis and table joining are overcome with help of patterns. There are various frameworks for handling multimethod, multifeature and multisource related issues. These frameworks are extracted from related business projects like different domains of insurance, banking, capital markets and government service. This has shown that the proposed frameworks are customizable and flexible for handling a large amount of the complex data. It involves multiple sources, methods and features as needed, for which data sampling and table joining may not be acceptable. It has also shown that the identified combined patterns are more actionable and informative than any single patterns identified in the traditional way.

References

[1] Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE Transactions on systems, man and cybernetics, VOL. 41, No. 3, June 2011

[2] Zhao, Y., Zhang, C., and Cao, L. (eds.) *Post-Mining of Association Rules:Techniques for Effective Knowledge Extraction*. Information Science Reference,2009.

[3] Ras, Z.W. and Wieczorkowska, A. Action-rules: how to increase profit of a company, in *Proceedings of PKDD'00*, LNAI 1910, 587-59, Springer, 2000.

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

[5] Zhao, and C. Zhang, "Mining impact-targeted activity patterns in imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 8, pp. 1053–1066, Aug. 2008.

[6] Cao, L., Yu, P.S., Zhao, Y. and Zhang, C. *Domain Driven Data Mining*, Springer, 2010.

[7] Cao, L., Luo, D. and Zhang, C. Knowledge actionability: satisfying technical and business interestingness, *International*

[8] *Journal of Business Intelligence and Data Mining*, 2(4): 496-514, 2007.

[9] Liu, B., Hsu, W. and Ma, Y. Pruning and summarizing the discovered associations. In *KDD99*, 125-134, 1999.

[10] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Combined pattern mining: From learned rules to actionable knowledge," in *Proc. AI*, 2008, pp. 393–403.

[11] Plasse, M., Niang, N., Saporta, G., Villeminot, A. and Leblond, L. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis 52*, 596-613, 2007.

[12] Ras, Z. and Dardzinska, A. Action Rules Discovery Based on Tree Classifiers and Meta-actions, *ISMIS09*, 66-75, 2009.

[13] Ozgur, A. Tan, P.-N., and Kumar, V. Rba: an integrated framework for regression based on association rules. *SDM'04*, 210-221, 2004.

[14] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Mining sequential patterns by pattern-growth: The PrefixSpan approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1424–1440, Nov. 2004.

[15] Wang, K., Jiang, Y. and Tuzhilin, A. Mining Actionable Patterns by Role Models, *ICDE06*, 16-25, 2006.

[16] Lesh, N., Zaki, M. J. and Ogihara, M. Mining features for sequence classification. In *KDD99*, 342-346, 1999.