

Informative Knowledge Discovery using Multiple Data Sources, Multiple Features and Multiple Data Mining Techniques

P.Sridevi¹, N.Venkata Subba Reddy²

¹Department of CSE, Sreenidhi Institute of Science & Technology, Ranga Reddy, Andhra Pradesh, India
Associate Professor

²Department of CSE, Sreenidhi Institute of Science & Technology, Ranga Reddy, Andhra Pradesh, India

Abstract: Data mining is a process of obtaining trends or patterns in historical data. Such trends form business intelligence that in turn leads to taking well informed decisions. However, data mining with a single technique does not yield actionable knowledge. This is because enterprises have huge databases and heterogeneous in nature. They also have complex data and mining such data needs multi-step mining instead of single step mining. When multiple approaches are involved, they provide business intelligence in all aspects. That kind of information can lead to actionable knowledge. Recently data mining has got tremendous usage in the real world. The drawback of existing approaches is that insufficient business intelligence in case of huge enterprises. This paper presents the combination of existing works and algorithms. We work on multiple data sources, multiple methods and multiple features. The combined patterns thus obtained from complex business data provide actionable knowledge. A prototype application has been built to test the efficiency of the proposed framework which combines multiple data sources, multiple methods and multiple features in mining process. The empirical results revealed that the proposed approach is effective and can be used in the real world.

Index Terms: Data mining, actionable knowledge discovery, multi-method mining, multi-feature mining, multi-source mining

I. INTRODUCTION

Data mining at enterprise level operates on huge amount of data such as government transactions, banks, insurance companies and so on. Inevitably, these businesses produce complex data that might be distributed in nature. When mining is made on such data with a single-step, it produces business intelligence as a particular aspect. However, this is not sufficient in enterprise where different aspects and standpoints are to be considered before taking business decisions. It is required that the enterprises perform mining based on multiple features, data sources and methods. This is known as combined mining. The combined mining can produce patterns that reflect all aspects of the enterprise. Thus the derived intelligence can be used to take business decisions that lead to profits. This kind of knowledge is known as actionable knowledge. The actionable knowledge is discovered through multiple data sources, multiple methods and multiple features. The intelligence thus obtained is dependable and reliable. As businesses are growing in complex data, there is a need for mining on complex data. However, it is challenging to discover actionable knowledge using complex data sources. This is because the generated business intelligence must be comprehensive and information that provides enough knowledge to take enterprise business decisions. There are many traditional methods being used in data mining. Therefore it is very challenging to combine them and use them to generate actionable knowledge. The challenges in the multi mining process can be categorized into multiple data sources, multiple methods, multiple features, post analysis and mining, joining multiple relational database and data sampling as well. Data sampling is generally not acceptable in real world data mining applications. Due to space and time limits combining multiple tables or joining them may not be possible. As data mining methods are developed for various data sources keeping some assumptions in mind, it is challenging to combine them for discovering actionable knowledge in real time applications. Combined association rules, combined rule clusters and combined rule pairs concepts are proposed in [1], [2] and [3]. These papers exposed the mining on complex data with respect to multiple data sets and obtaining comprehensive knowledge. Combined association rule is nothing but a set of multiple item sets that are heterogeneous in nature. Combined rules pairs are nothing but derived from combined association rules by combining rule clusters. The combined rule pairs are also derived from combined rules only. Traditional algorithms such as FPGrowth [4] can't be used to derive combined association rules. This paper aims at presenting a new comprehensive framework for combined mining. As a matter of fact, this paper makes use of existing methods or techniques as part of the framework. Therefore it integrates multi source combined mining, multi-method combined mining, and multi-feature combined mining. Multiple features might include demographics of customer, behavior, business impacts and also transactional data. Multi method might include clustering, classification and so on. Multiple data sources do mean that the mining process takes data

from multiple related data sources. The deliverables of the proposed framework combined patterns or combined association rules. The following are the general aspects of combined mining.

- Combined patterns generated by using various features can reflect the characteristics of enterprise closely and such business intelligence is always reliable.
- The combined mining with respect to multiple data sources can give patterns that cover many aspects of the enterprise as data is taken from different data sources.
- Multiple methods mining results in patterns that reflect the in depth nature of data and also the advantages provided by various mining algorithms are with this kind of combined mining.
- Metrics are of multiple interestingness in nature can be applied to generated patterns that can verify the significance of generated patterns. Instead of providing a specific algorithm and get results pertaining to that algorithm, it is good idea to combine them and get more accurate and actionable knowledge.

The contributions of this paper include the concept of combined mining that is based on the existing works. It discusses the framework and various combined mining such as multi-source combined mining, multi-method combined mining, and multi-feature combined mining. Provides various techniques for pattern interaction and novel patterns such as clustered patterns, combined patterns etc. Interestingness metrics are evaluated. Discovering practically the combined patterns or actionable knowledge from real world businesses such as banks.

II. RELATED WORK

Mining is a process of extracting trends or patterns from historical data. These trends or patterns can provide business intelligence that leads to actionable knowledge. There are many data mining methods or algorithms that exist for mining data to get patterns. However, all the existing algorithms are single-step mining algorithms. This means that they provide business intelligence inadequately. They may not be able to reflect the complex needs of an enterprise to take decisions correctly. When multiple data mining techniques are combined it is possible to get actionable knowledge that can cater to the needs of an enterprise. In this paper combining mining algorithms [1] have been implemented using a prototype application that demonstrates the efficiency of combined mining. The combined actionable knowledge can't be provided by existing algorithms such as FPGrowth [4]. The existing works on data mining operations on complex data or enterprise generally of different types such as direct mining approaches; post mining of patterns; data sets with extra features; multiple methods integration; and also joining multiple relational tables. Harmony [5] proposed an approach to mine for discriminative patterns. Other such experiments include contrast patterns [6], model based search tree [7]. These algorithms attempted to use multiple features or mining techniques. Combined mining is best used to provide actionable knowledge in spite of complex data sets, and features.

There are four categories of combined mining approaches in literature. A commonly used approach [8] is the post mining or post analysis of obtained patterns. It is best used to prune the rules obtained after mining database or reducing redundancy or even summarizing the patterns obtained [9]. In [1] combined mining proposed contain direct mining methods. In [1] multisource combined mining, multi-method combined mining and multi-feature combined mining were introduced. These algorithms are used in this paper and implemented in the prototype application that is used to demonstrate the efficiency of the combined mining. In these approaches also post mining of patterns is used when it is necessary. Multisource combined mining considers data of various natures. It does mean that it combines multiple data sets as it is required by an enterprise which has data of different kinds. The result of multisource data mining is the actionable knowledge that reflects different angles of an enterprise.

Multi-feature data mining is used to have data with multiple features. This kind of mining also leads to actionable knowledge that reflects the complex needs of an enterprise. Here multiple features are used. For instance the prototype application in this paper has been implemented using banking data containing customers' demographic data and transactional data. Multi-method combined mining mixes up many existing data mining techniques that are very useful independently for a particular purpose only. However, in multi-method combined mining various mining techniques are combined to obtain actionable knowledge. For instance apriori and ID3 can be combined. The apriori algorithm provides association rules while the ID3 is meant for providing decision tree. The combination of these two can provide actionable knowledge that helps in taking well informed decisions. Other data mining algorithms that are existed and can be used in combination are clustering, rarity mining, regression, association rule mining, sequence classification [15], rule based mining [10] etc. With the help of combined mining sequential and cluster patterns can be used further. It does mean that the result of one method can be used in another method and the process can be repeated if required. It is also possible to have a chain of mining operations until the desired actionable knowledge is derived.

III. PROPOSED MINING FRAMEWORK

The proposed framework combines the existing algorithms or methods to integrate multiple features, multiple data sources and multiple mining methods like classification. The combined mining can be done with multiple features, multiple data sources and multiple methods. The architecture of the proposed framework is shown in fig. 1 which is common framework for all these combined mining approaches.

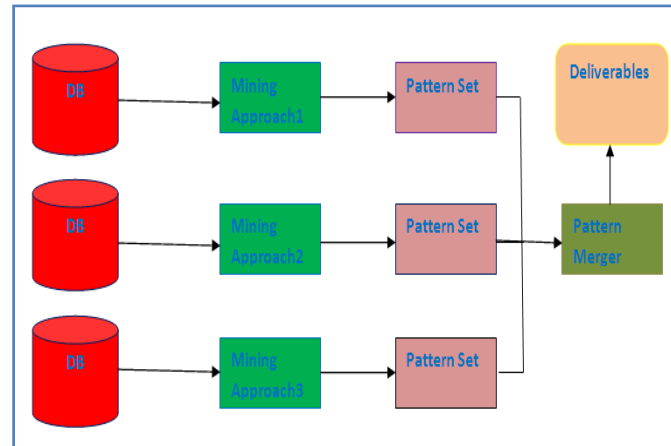


Fig. 1 – Architecture of proposed mining framework [1]

The common architecture for multi-feature mining, multi-method mining and multi-source mining are described here. There are multiple mining approaches that make use of given data source (s) and perform one or more mining algorithms or methods. Each mining approach produces a set of patterns. All the patterns obtained from multiple mining approaches are merged together. This is done by a component known as pattern merger. Pattern merger is a program that combines all pattern sets obtained from various mining techniques. The result of pattern merging process is the final deliverables. The deliverables are nothing but actionable knowledge that facilitates comprehensive decision making. This framework is useful to all enterprises where there is complexity in business data and business intelligence has to cover many aspects of that business. Especially this framework is suitable for domains like banking, insurance where monetary transactions are stored and maintained. The historical data has to be mined and right decisions are to be made in case of all monetary decisions such as issuing loans and other customer centric services. The algorithms of [1] are used for the experiments made using a prototype application. The ensuing sections present algorithms and the experimental results.

IV. ALGORITHMS

This section provides description and algorithms for multisource combined mining, multi-feature combined mining and multi-method combined mining.

Multisource Combined Mining

The combination of multiple data sources (D): The combined pattern set P consists of multiple atomic patterns identified in several data sources. By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.

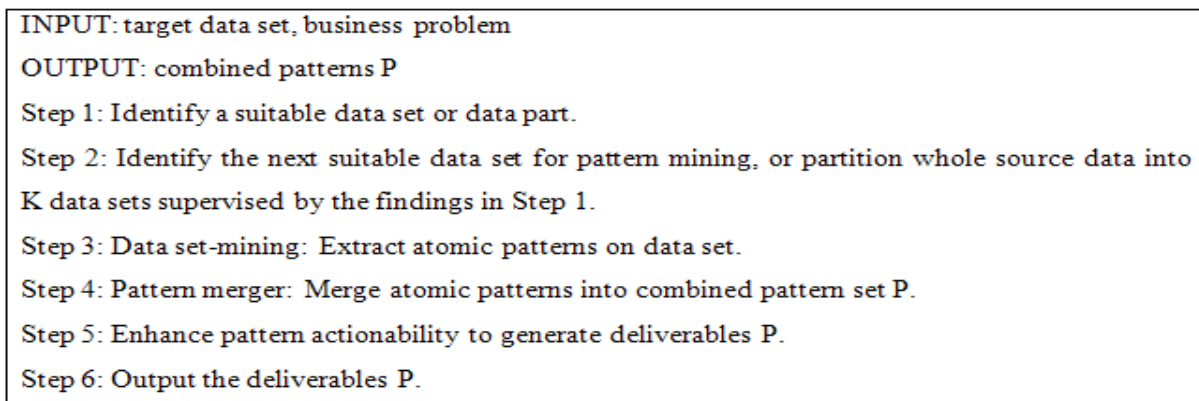


Fig. 2 – Algorithm for multisource combined mining

Multi-feature Combined Mining

The combination of multiple features (F): The combined pattern set involves multiple features, namely, e.g., features of customer demographics and behavior. By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.

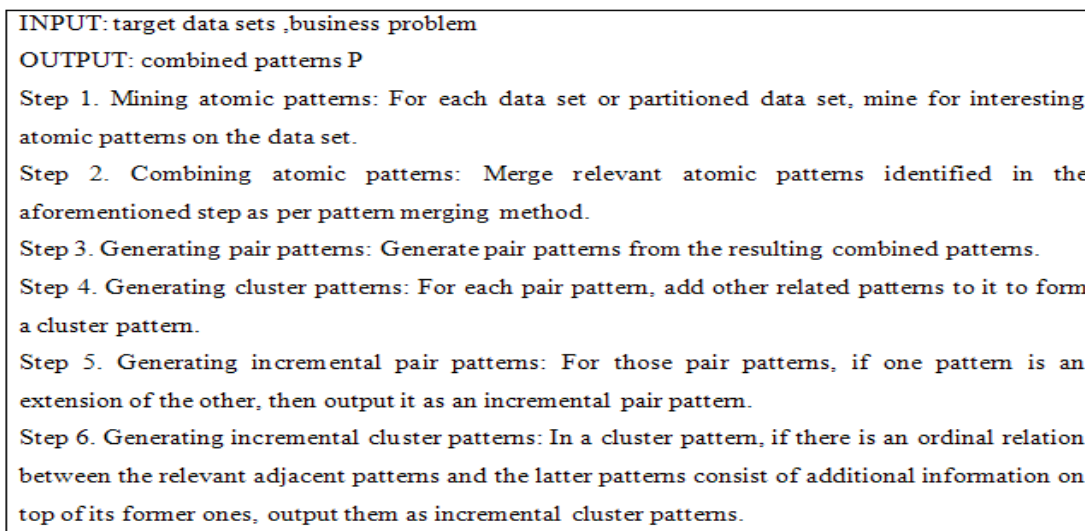


Fig. 3 – Algorithm for multi-feature combined mining

Multi-method Combined Mining

Multi-method combined mining is another approach to discover more informative knowledge in complex data. The focus of multi-method combined mining is on combining multiple data mining algorithms as needed in order to generate more informative knowledge. The combination of multiple methods (R): The patterns in the combined set reflect the results mined by multiple data mining methods like association mining and classification. 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. The algorithms used for multi-method combined mining are Apriori Algorithm and ID3 Algorithm. These two algorithms are well known and existing algorithms in data mining domain.

V. EXPERIMENTAL RESULTS

The environment used to built the prototype application used to demonstrate the efficiency of various combined mining algorithms described in the previous section is Java Standard Edition (JSE 6.0), Net Beans IDE and Oracle 10G Express Edition. The hardware used is a PC with 2GB RAM, and 2.9x GHz processor. The dataset used for the experiments is pertaining to banking domain.

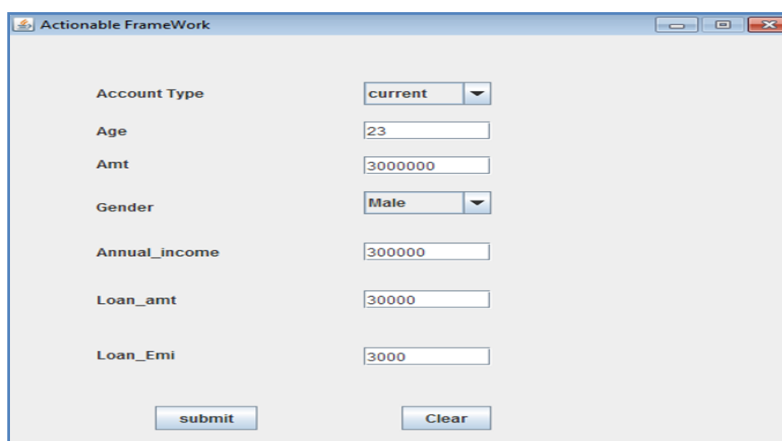


Fig. 4 – Input Screen for synthetic dataset

The dataset used in related to banking domain. The data mining makes use of customers' demographic and also transactional data in order to find patterns and provide actionable knowledge finally.

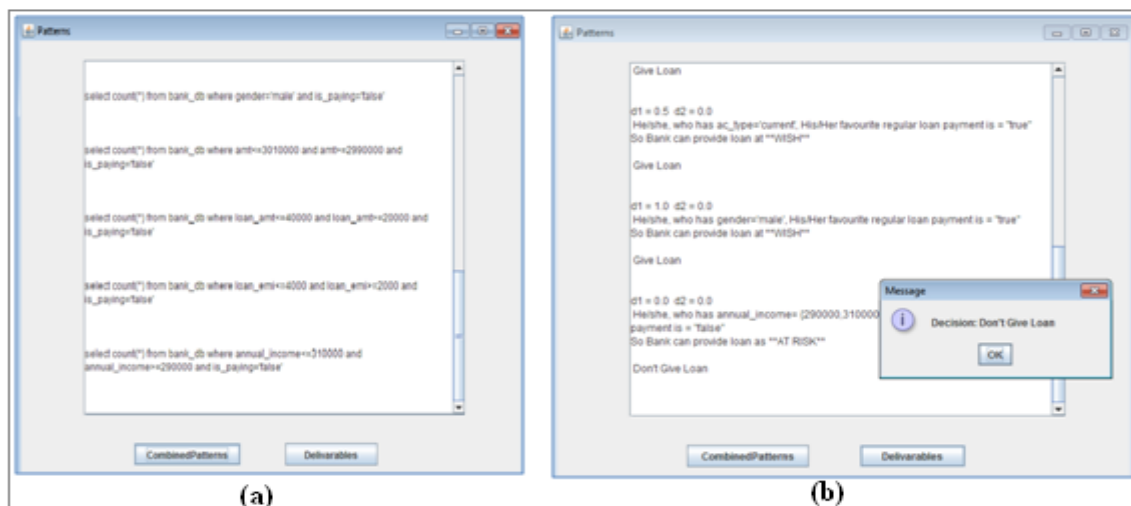


Fig. 5 – Results of multisource combined mining

As can be seen in fig. 5, (a) shows patterns while (b) shows actionable knowledge that is whether loan can be given to given customer or not.

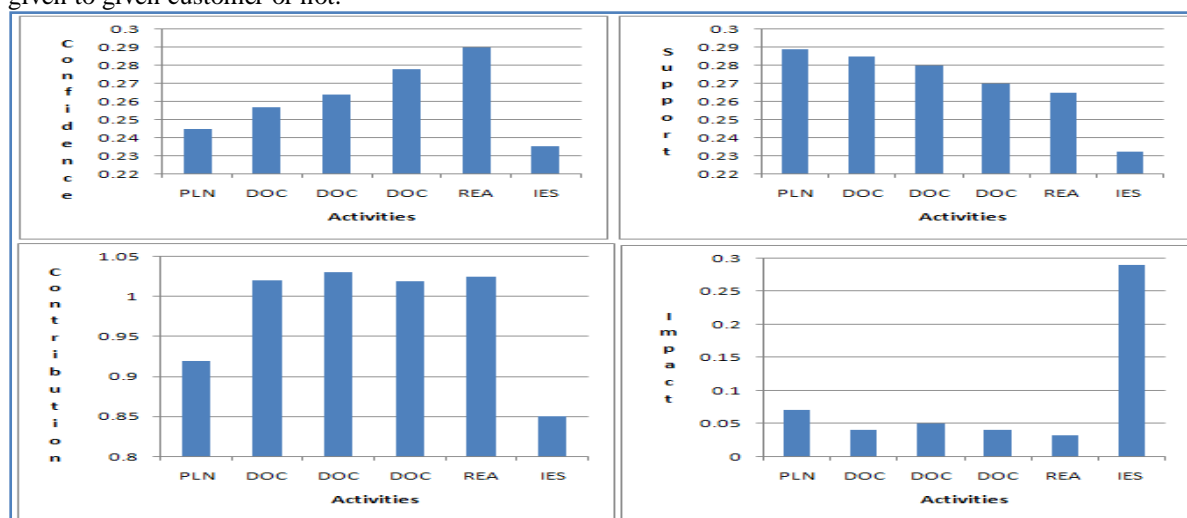


Fig. 6 –Dynamics of Incremental Cluster Patterns

As can be seen in fig. 6, it visualizes the dynamics of clusters patterns in terms of confidence, support, contribution and impact. The PLN, DOC, DOC, REA, IES represent a series of activities.



VI. CONCLUSION

Enterprises generally are distributed in nature with multiple databases, multiple servers and so on. The data in such businesses is very complex. Often the applications exhibit multiple features and the applications also heterogeneous in nature. For instance the database might have details pertaining to business impact, business appearance, service usage, behavior, preferences and demographics. Considering complex business scenarios and requirements of business intelligence, it is time consuming to have multiple single mining methods or features or data sources. The single piece of information from a single mining approach may not reflect actual requirement of the enterprise. Therefore, we proposed a mining framework that combines multiple data sources, multiple methods and multiple in order to obtain multiple pattern sets and finally the deliverables that is actionable knowledge that helps in taking comprehensive and well informed decisions. Such combined mining frameworks yield best actionable knowledge that can provide profits to organizations that use it. Thus the proposed framework can be used in domains like banking, insurance etc. to take effective decisions in monetary matters. Multiple data sources, features, and methods used in the proposed framework are not entirely new things. They are taken from existing works and customized to get the required framework for combined mining. The data mining framework is applied to many real time solutions and the experiments made with prototype application revealed that the framework is very effective and able to provide actionable knowledge.

REFERENCES

- [1] Longbing Cao, Senior Member, IEEE, Huaifeng Zhang, Member, IEEE, Yanchang Zhao, Member, IEEE, Dan Luo, and Chengqi Zhang, Senior Member, IEEE. Combined Mining: Discovering Informative Knowledge in Complex Data. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 41, NO. 3, JUNE 2011
- [2] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, “Combined association rule mining,” in Proc. PAKDD, 2008, pp. 1069–1074.
- [3] 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.
- [4] 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.
- [5] J. Wang and G. Karypis, “HARMONY: Efficiently mining the best rules for classification,” in Proc. SDM, 2005, pp. 205–216.
- [6] G. Dong and J. Li, “Efficient mining of emerging patterns: Discovering trends and differences,” in Proc. KDD, 1999, pp. 43–52.
- [7] W. Fan, K. Zhang, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure, “Direct mining of discriminative and essential graphical and itemset features via model-based search tree,” in Proc. KDD, 2008, pp. 230–238.
- [8] Y. Zhao, C. Zhang, and L. Cao, Eds., *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. Hershey, PA: Inf. Sci. Ref., 2009.
- [9] B. Liu, W. Hsu, and Y. Ma, “Pruning and summarizing the discovered associations,” in Proc. KDD, 1999, pp. 125–134.
- [10] K. K. R. Hewawasam, K. Premaratne, and M.-L. Shyu, “Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 6, pp. 1446–1459, Dec. 2007.

AUTHORS

	P. Sridevi is a student of Sreenidhi Institute of Science and Technology, Ranga Reddy, Andhra Pradesh, India. She has received B.Tech degree in Computer Science and Engineering and M.Tech Degree in Software Engineering. Her main research interest includes Data Mining and Software Engineering.
	N. Venkata Subba Reddy is working as Associate Professor at Sreenidhi Institute of Science & Technology, Ranga Reddy, and Andhra Pradesh, India. He has received M.Tech Degree in Software Engineering. His Main Interest includes Data Mining and Web Technologies.