ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

Discovering Informative Knowledge in Complex Data Using Pre And Post Process Mining

S.ArunKumar,S.SundarRajan,

PG scholar, Department of Computer Science and Engineering, Surya Group of Institutions, Vikiravandi, Villupuram District, India

Associate Professor, Department of Computer Science and Engineering, Surya Group of Institutions, Vikiravandi, Villupuram District, India.

ABSTRACT:Over last 2 decade data mining practices have actively employed in various fields. Input data for such indiscriminate applications not only varies but is also complex i.e. large as well as diversified in nature Many data mining applications are limited to mine only a certain type of data. Moreover, these applications use solitary technique of data mining to discover knowledge. They provide business intelligence in all aspects. That kind of information can produce the actionable knowledge. Recently data mining has got tremendous usage in the real world Combined mining which is one of the novel approach to mine complex data. It embraces three different framework viz. multi-source combined mining, multi-method combined mining, and multi-feature combined mining. The outcome of combined mining process is named as combined patterns, which are indisputably more informative than simple association rules. Domain knowledge plays major role in data mining applications. Out of various ways of representing domain knowledge, ontology is effective one. The combined patterns can be converted into more practicable information if one processes them with the help of domain knowledge. The purpose of this work is to study the impact of incorporating domain knowledge in pre-post process mining of combined patterns. So we will initially formalize user's knowledge & goals using rules, schema and then filter the extracted combined patterns using various operators.

KEYWORDS: Combined Mining, Multi Source Combined Mining, Multi Method Combined Mining, Multi Feature Combined Mining, Ontology, Pre Processing ,Post Processing , Rule schema, Operators..

OBJECTIVES

- To gain an understanding why the Combined mining is a novel approach to mine complex data .
- To identify key issues relevant to combined.
- Mining methods and pre-post processing mining

I. INTRODUCTION

Data mining by now is being widely used in many areas such as public services, telecom, share market, shopping malls, health care and many more. In the state of art , developers also mind about the various types of data sources used inthe applications. These days, data sources entail heterogeneous data for example transactional data, XML documents, text files etc. Also the transactional data sources may hold multiple features. To handle such multi-featured and heterogeneous data sources, the process of mining needs to be generalized. Combined mining will provide an overall solution to face the challenge of mining complex knowledge in complex data

We mentioned earlier discussions show the need or developing capable techniques for involving multiple heterogeneous features, data sets, and methods in enterprise data mining. As we will discuss in Section II about related works, the existing works in handling the aforementioned challenges can be categorized into the following

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

aspects: 1) data sampling; 2) joining multiple relational tables; 3) post analysis and mining; 4) involving multiple methods; and 5) mining multiple data sources. In real-life data mining, data sampling is often not acceptable since it may miss important data that are filtered out. Table joining may not be possible because of time and space limit such as in dealing with hundreds of millions of transactions from multiple sources in our case studies. In addition, techniques for involving multiple methods and handling multiple data sources are often specifically developed for particular cases

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 produce which reflect multiple aspects of nature 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.

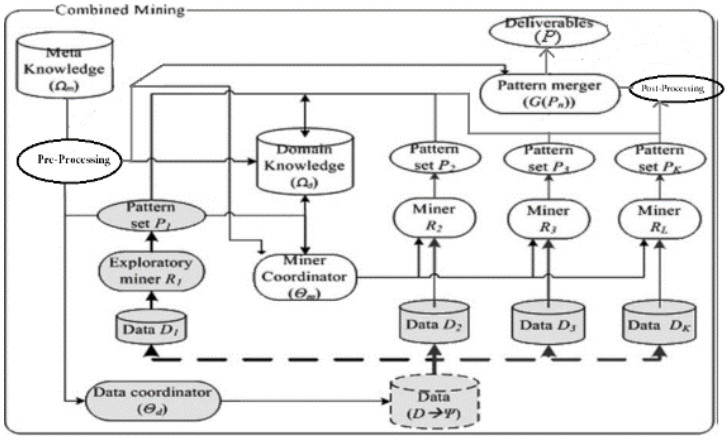


Fig. 1:Combined mining for actionable patterns

Instead of presenting a specific algorithm for mining a particular type of combined patterns, this paper eye on abstracting several common and flexible frameworks from the structure perspective, which can foster wide implications and particularly can be instantiated into many specific methods and algorithms to mine for various patterns in complex data

II. LITERATURE REVIEW

Combined mining concept introduced in [1] enforce to make use of more than one mining techniques at the same time. Integrated use of classification and association rule mining was also done. The integration of classification and done by mining a special subset of association rules, i.e. CARs. Classifier built is more accurate than that produced by the simple classification system

To work with complex, huge, and heterogeneous data many efforts has been taken. Association rules are generally extracted from transactional data with a single set. In [2], a novel approach for extracting combined association rules was proposed. Combined association rules were prearranged as rule sets, each of which is composed of a number of single combined association rules. To achieve this, association rule mining was done in two steps, 1) rule generation and 2) definition of new interestingness measures. In rule generation, the frequent item-sets were discovered among item-set groups to improve efficiency. Then new interestingness measures were defined to discover more actionable knowledge This project uses the concepts of combined association rules, combined rule pairs, and combined rule clusters which are described in [3]. The concepts were proposed to mine actionable patterns from data. Also combined pattern mining is extended to use complex data i.e. multiple, heterogeneous data sets in [1].

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

Second, approaches to mining for more informative and actionable knowledge in complex data can be generally categorized as follows: 1) direct mining by inventing effective approaches; 2) post-analysis and post-mining of learned patterns; 3) involving extra features from other data sets; 4) integrating multiple methods; and 5) joining multiple relational tables.

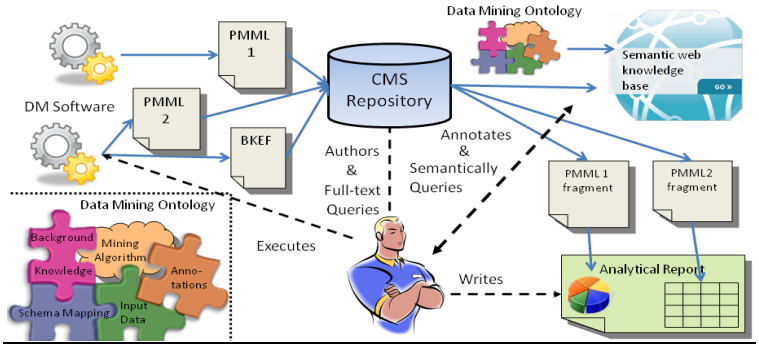


Fig 2:Ontology process

The post analysis and post mining of learned patterns is a commonly used approach ,for instance, to prune rules, reduce redundancy [10], and summarize learned rules [12]. Different from post analysis-based methods, most of the combined patterns introduced in this paper can be generated directly.

Some of them can be identified through the post analysis. Aside from the direct mining of combined patterns, the post mining of the identified patterns can be conducted where necessary in order to make the patterns more actionable. For example, the multi feature combined mining approach considers features from multiple data sets during the direct generation of more informative patterns

III. ONTOLOGY FOR DOMAIN KNOWLEDGE

There are many tools and forms available for representation of knowledge. Knowledge specification language is used to build “is-a organization” of database attributes. This type of organization is called as “Item taxonomy”. Practically, user might want to use concepts that are more expressive and accurate than generalized concepts and that result from relationships other than the “is-a” relation. Ontology helps here. It includes the features of taxonomies but adds more representation power. So we can say that, taxonomy is simply a hierarchical categorization or classification of items in a domain. On the contrary, an ontology is a specification of several characteristics of a domain, defined using an open vocabulary.

A. leaf concepts

A leaf concept is defined such that, each leaf concept is associated to one item in the database.

B. Generalized concepts

Generalized concepts are defined such that the concepts subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts.

C. Restricted concepts

Restriction concepts are described using logical expressions defined over items and depend on the user individually.

IV. DATA PRE-PROCESSING

The first phase of this paper discusses on data pre-processing algorithms used to clean raw log data. The purpose of data pre-processing is to extract useful data from raw web log and then transform these data in to the form necessary for pattern discovery. Due to large amount of irrelevant information in the web log, the original log cannot be directly used in the web log mining procedure, hence in data pre processing phase, raw Web logs need to be cleaned, analyzed

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

and converted for further step. The data recorded in server logs, such as the user IP address, browser, viewing time, etc, are available to identify users and sessions.

- Here we using HB-reduction algorithm for pre-processing to clean the data Horizontal reduction is performed on the data set by examining attributes one by one
- A lot of computational intensive operations in our algorithms are performed using the database *Count, Update operations.*
- DB-Hreduction algorithm reduces both the attributes and tuples of the data set
- Which reduces the search spaces to maximum extent without losing essential information

However, because some page views may beached by the user browser or by a proxy server, we should know that the data collected by server logs are not entirely reliable. This problem can be partly solved by using some other kinds of usage information such as cookies cleaning the data, we can create the database according to our application which includes the information about user identification, session identification, path completion etc. By using the same proxy server, different users leave the same IP address in the server log, which makes the user identification rather difficult. [6] presented the solution to these problems by using Cookies or Remote Agents. In our work the various processes involved in data pre-processing. This module includes the identification of users and user's sessions, which are used as basic building blocks for pattern discovery.

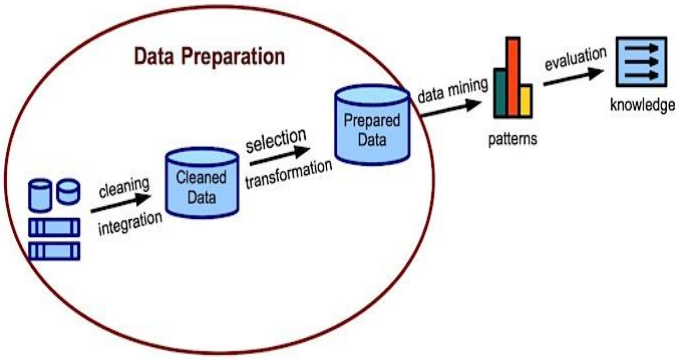


Fig.3: Pre-processing process

V. DATA POST PROCESSING

It is g as an important component of KDD consists of many various procedures and methods that can be categorized into the following groups.

- ***Knowledge filtering:*** *Rule truncation and post pruning.* If the training data is noisy then the inductive algorithm generates leaves of a decision tree or decision rules that cover a very small number of training objects. This happens because the inductive (learning) algorithm tries to split subsets of training objects to even smaller subsets that would be genuinely consistent. To overcome this problem a tree or a decision set of rules must be shrunk, by either post pruning (decision trees) or truncation (decision rules)

- ***Interpretation and explanation****:* Now, we may use the ac-quired knowledge directly for prediction or in an expert system shell as a knowledge base. If the knowledge discovery process is performed for an end-user, we usually document the derived results. Another possibility is to visualize the knowledge [9], or to transform it to an understandable form for the user-end. Also, we may check the new knowledge for potential conflicts with previously induced knowledge. In this step, we can also summarize the rules and combine them with a domain-specific knowledge pro-vided for the given task.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

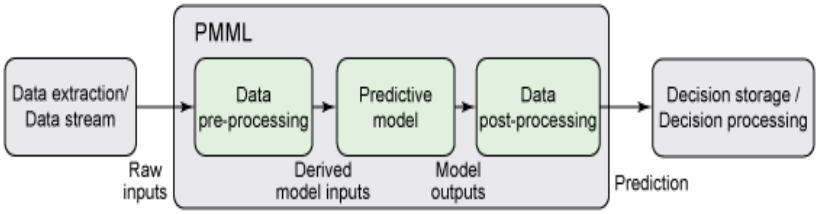


Fig.4: Post processing process

❖ **Evaluation:** After a learning system induces concept hypothe-ses (models) from the training set, their evaluation (or testing) should take place. There are several widely used criteria for this purpose: classification accuracy, comprehensibility, computational complexity, and so on.

❖ **Knowledge integration*:*** The traditional decision-making systems have been dependant on a single technique, strategy, model. New sophisticated decision-supporting systems combine or refine results obtained from several models, produced usually by different methods. This process increases accuracy and the likelihood of success.

- For post processing here we use csDecision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree technique used for classification of a dataset.

- They provide a set of rules that can apply to a new dataset to predict which records will have a given ending
- CART segment a dataset by create two way split while CHAID segment using chi square tests to create multi-way splits. CART classically requires fewer data training than CHAID

VI. APRIORI ALGORITHM

Figure 1 gives the Apriori algorithm. The rst pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A subsequent pass, say pass k, consists of two phases. First, the large itemsets $L^{k,}1$ found in the (k,1)th pass are used to generate the candidate itemsets C^k , using the apriori-gen function described in Section 2.1.1. Next, the database is scanned and the support of candidates in C^k is counted.

L^1 = flarge 1-itemsetsg;
1. for (k = 2; $L^{k,1}$ 6= ;; k++) do begin
2. C^k = apriori-gen($L^{k,1}$); // New candidates
3. Forall transactions t 2 D do begin
4. C^t = subset(C^k , t); // Candidates contained in t
5. Forall candidates c 2 C^t do
6. c:count++;
7. end
8. L^k = fc 2 C^k j c:count minsupg
9. end
10. Answer = $_S$kL^k ;

Figure 1: Algorithm Apriori

VII.CONCEPT OF COMBINED MINING

A. Basic concepts

We introduce the concept of combined (pattern) mining. Combined mining is introduced for handling the complexity of employing multi-feature sets,multi-information sources,constraints,multi-methods and multimodels in datamining,and for analysing complex relations between objects or descriptors or between identified patterns during the learning process. Combined patterns maybe formed by t h e analysis of the internal relations between objector pattern constituents

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

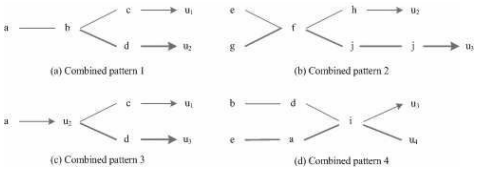
obtained by a single method on a single dataset, for instance, combined sequential patterns formed from analysing the relations within a discovered sequential pattern space

With the enforcement of object and pattern relation analysis, which is a very new topic in the data mining community, many approaches and algorithms are available in the literature on other aspects of the above combinations. The main involvement of combined mining is that it enables the extraction, discovery, construction and induction of knowledge which consists not simply of discriminant objects but also of communication and relations between the objects, as well as their impact. This is called actionable complex patterns, because they send back pattern elements and relations, which form certain pattern structures and dynamics, and indicate decision-making actions.

Combined mining gives an overall solution for facing the challenge of mining complex knowledge in complex data , It also substantially develop upon other individual methods such as conceptual inductive learning and inference, generalization, aggregation and summarization in order to combine together with data-driven knowledge discovery from complex environments. Specifically, pattern relation analysis augments the following areas: knowledge representation and reasoning, inductive learning, semantic and ontological engineering, pattern theory, and pattern language.

B.Combined Pattern Paradigms

The combination of the aforementioned pattern elements and combination factors in terms of particular selection criteria and pattern coupling relations will con- tribute to different pattern paradigms.



1) Figure4: ComplexCombinedPatterns

Below, we discuss three combined pattern paradigms: similar combined patterns, dissimilar combined patterns, and dependent combined patterns.

Scenario 1 (Similar Combined Patterns) The constituent patterns in a combined pattern share some similarity in feature, interaction, relation, structure or impact. As discussed above, similarity is measured through distance, density, shape, structure or relation that is specific to certain pattern mining methods. Typical similar combined patterns include the combinations of frequent patterns, high utility patterns, clusters, and classes.

Scenario 2 (Dissimilar Combined Patterns) The constituent patterns in a combined pattern have some dissimilarity in feature, interaction, relation, structure or impact, as measured by distance, density, shape, structure or relation specific to certain pattern mining methods. Typical dissimilar combined patterns are contrast-based combined patterns, in which two atomic patterns are associated with opposite impacts (labels).

Scenario 3 (Dependent Combined Patterns) Also called Conditional Combined Patterns, in which on pattern forms the precondition of another. The condition may come from the aspect of feature, interaction, relation, structure or impact. Three types of dependent combined patterns are incremental patterns, decremental patterns, and conditional probability patterns.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

Instance 1 (Incremental Patterns) Also called Prefix Combined Patterns, in which any two neighbouring atomic patterns in the combined collection form an incremental relation, namely pattern i + 1 sharing some incremental part of features, pattern elements, structures or impacts on top of pattern i.

Instance 2 (Decremental Patterns) Also called Postfix Combined Patterns, where the constituent pattern i consists of an additional part of features, pattern elements, structures or impacts compared to pattern i + 1.

In incremental and decremental patterns, some atomic patterns serve as the underlying patterns; the immediate neighbouring patterns are the derivative form of them.

Incremental and decremental patterns are more about pattern structure dependence. Readers may refer to to access details about how incremental and decremental frequent patterns and frequent sequences are formed in social security data. In practice, such structural relations are often hard to detect in large data; this is even challenging which structural relations appear to be implicit. This is particularly difficult if we do not have the hypothesis of what kind of incremental or decremental relations exist in the data. For patterns appearing in trees, graphs or other unstructured formats, it would be much more difficult to learn and extract such relations.

In addition, conditional probability patterns cater for implicit dependency be- tween atomic patterns when a probabilistic model fits a dataset.

Instance 3 (Conditional Probability Patterns) The constituent patterns form a conditional probability relation in terms of features, elements, interaction, structure or impact.

The conditional probability relation may be embodied through certain statistical relations and functions. For instance, a chain of states may affect one another, which can be modelled according to the Markov assumption.

VIII. CONCLUSION

In This paper has presented a comprehensive and general approach named combined mining for discovering informative knowledge in complex data. We focus on discussing the frameworks for handling multi-feature, multisource, and multi-method related issues. We have ad-dressed challenging problems in combined mining and summarized and proposed effective pattern merging and inter-action paradigms, combined pattern types, such as pair patterns and cluster patterns, interestingness measures, and an effective tool dynamic chart for presenting complex patterns in a business-friendly manner using pre-processing and post-processing methods on ontology concepts.

The frameworks are carry out from our relevant business projects conducted and currently under investigation from the domains of government service, banking, insurance, and capital markets, financial, shopping etc...because the usage of data mining spread all over business places need of discover knowledge is increase day by day , Several real-life cases studies have been briefed which instantiate some of the proposed frameworks in identifying combined patterns in multiple sources of governmental service data. 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.

We are further developing effective paradigms, combined pattern types, combined mining methods, pattern merging methods, it may be vary from one industry project to other choose effective methods or patterns depends project we select and interestingness measures for handling large and multiple sources of data available in our industry projects

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 4 , April 2015

for government, stock market, insurance, and banking, in future the data mining play the vital role in our daily life is possible because of monster development data usage in various fields in all applications so we need to develop the effective patterns to discover the knowledge from complex day

REFERENCES

[1] L. Cao, Y. 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.
[2] L. Cao, Y. Zhao, H. Zhang, D. Luo, and C. Zhang, "Flexible frameworks for actionable knowledge discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1299–1312, Sep. 2010.
[3] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective lassification," in *Proc. ICDE*, 2007, pp,716-725.
[4] S. Dzeroski, "Multirelational data mining: An introduction," *ACMSIGKDD Explor. Newslett.*, vol. 5, no. 1, pp. 1–16, Jul. 2003.
[5] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proc. KDD*, 1999, pp. 43–52.
[6] W. Fan, K. Zhang, J. Gao, X. Yan, J. Han, P. Yu, and O. Verscheure, "Di-rect mining of discriminative and essential graphical and itemset features via model-based search tree," in *Proc. KDD*, 2008, pp. 230–238.
[7] K. K. R. Hewawasam, K. Premaratne, and M.-L. Shyu, "Rule mining and classification in a situation ssessment application: A belief-theoretic ap-proach for handling data imperfections," *IEEE Trans. Syst., Man, Cybern.B, Cybern.*, vol. 37, no. 6, pp. 1446–1459, Dec. 2007.
[8] Combined Mining: Analyzing Object and attern Relations for Discovering Actionable Complex PatternsLongbingCaoAdvanced AnalyticsInstitute,University of Technology Sydney,

AUTHORSBIOGRAPHY

S.ARUNKUMAR received the MSC software Engineering from Muthayammal college of Engineering which belongs to Anna University. He is currently pursuing his M.E degree in Computer Science and Engineering in Surya Group of Institutions which belongs to Anna University, His research interest includes Cloud Computing, Cloud related hypervisor technologies &Data Mining, Information & Knowledge Management .He has published his research work in cloud computing and complex data in a leading international journals.

S. SUNDAR RAJAN is an Associate Professor of Computer Science & Engineering at Surya Group of Institutions, Vikravandi. He is a research scholar at St. Peters University, Chennai. He received his Master degree from Anna University, Chennai. His research interest includes Data Mining, Information & Knowledge Management, Cloud Computing and Computer Communication & Networks. He has published his research work in cloud computing and complex data in a leading international journals