EFFICIENT MINING OF MULTI DIMENSIONAL DATA USING COMBINED PATTERNS

N.B.S. Vijay Kumar*

K.Kalyani, (M.Tech)**

Abstract

In data mining, Mining Informative Knowledge from heterogeneous data has been recognized as one of the most challenging problem. The proposed concept of combined mining used to extract actionable knowledge from complex data .The proposed concept is logically divided into three functional requirements namely multi source combined mining, multi feature combined mining and multi method combined mining. It combines the patterns either from the multiple data sources or from the multiple features data.Association-Classification methods are used in the multi mining method to mine the efficient patterns from the heterogeneous data. The Association algorithm namely Apriori algorithm is used to mine the frequent patterns from the data sources. After finding the frequent patterns from this method, again another method called classification algorithm namely ID3 is used to classify the efficient combined patterns from the heterogeneous data so that it will be easy for decision makers to take correct decisions for business impact. A bank application is taken as example and identified the efficient combined patterns for the clearance of the loan and improving the service objectives of the bank.

Keywords - Combined Patterns, Frequent Patterns, Data Bases, Iterative Dichotomiser 3, Class Association rules

* Department of CSE, GATES Institute of Technology

** Assistant professor, Department of CSE, ALITS Engineering college

1. Introduction

Enterprise data mining applications, such as mining public service data and telecom fraudulent activities, 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, business people certainly expect the discovered knowledge to present a full picture of business settings rather than one view based on a single source. Knowledge reflecting full business settings is more business friendly, comprehensive, and informative for business decision makers to accept the results and to take operable actions accordingly. With the accumulation of ubiquitous enterprise data, there is an increasing need to mine for such informative knowledge in complex data.

It is challenging to mine for comprehensive and informative knowledge in such complex data [6] suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, the traditional methods [3] 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.

1.2 Objective

The objective of this project is to find the efficient combined patterns in the heterogeneous data. Using Association-Classification algorithms the combined patterns are found in the complex data, so that it is easy for decision makers to take decisions for business impact.

2. Methodology

Mining of informative patterns from heterogeneous data can be done through the concept of combining mining.

2.1 Combined Mining

The combined mining approach is a two-to-multistep mining procedure, which mine the actionable patterns from complex data. Combining components from either multiple data sources or multiple features or by applying multiple methods on individual data sources. 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.

From the above the word Combined principally referred as:

- The combination of multiple data sources principally refers to the atomic patterns that are identified in the individual sources and finally combined to have the efficient patterns that lead to informative knowledge.
- The combination of multiple features principally refers to the heterogeneous data sources like customer demographic data, transactional data, time series data etc.
- The combination of multiple methods principally refers to individual methods are applied on individual data sources and finally these are combined to reflect the deliverables called combined patterns.

2.2. Multi Source Combined Mining

Real world enterprise applications often involve multiple heterogeneous and distributed data sets that cannot or are too costly to be integrated. Another common situation is where the data volume is so large that it cannot be handled by scanning the
whole data set.

Such data have to be partitioned into either small and manageable sets or in terms of business categories such as billing, networking, and accounting data in telecommunication systems. Mining such complex data requires the handling of multi data sources implicitly or explicitly.

Business data is often distributed amongst different databases, relational tables, files, systems and/or geographic locations. Mining this type of data structure, to extract business insight, is difficult and subject to ongoing research, because the existent algorithms work on de-normalized file structures, either on a single flat file.

Algorithm scalability issues concerned with computing time and memory space can be prohibitory expensive, also privacy and integrity issues play an important role.

Heterogeneous data sources, such as demographic and transactional data, are part of everyday business applications and used for data mining research. Traditional data mining algorithms are not applied, directly, to the above data structures. From a business perspective, patterns extracted from a single normalized table or subject file are less interesting or useful than a full set of Data sets.

The following figure1 illustrates the discovery of combined patterns either in multiple data sets or sub sets (Data set A, Data set B & Sub sets of Data set B).Using mining methods patterns are identified from multiple data sources these patterns are combined and finally actionable patterns are taken as deliverables.

Multiple patterns extracted from different datasets. Here combination of multiple data sources (*D)*: The combined pattern set *P* consists of multiple atomic patterns identified in several data sources. For example, in clearance of bank loan the following two data sets such as customer demographic data and transactional data are two data sets involved in mining for demographic–transactional patterns.

Customer ID	Gender	…………
1	M	
5	F	

Table 1: Transactional Demographic Data.

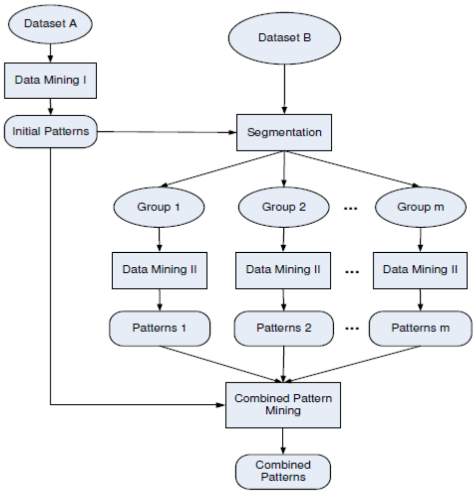

Figure 1: Combined Patterns from Multiple

2.3Multi Feature Combined Mining

In multi feature combined pattern (MFCP) mining, a combined pattern is composed of heterogeneous features of different data types, such as binary, categorical, ordinal, and numerical, or of different data categories, such as customer demographics, transactions, and time series.

Customer ID	Polices	Activities
1	(c1,c2)	(a1-a2)
4	(c2,c4)	(a1-a2-a5)
5	(c1,c2,c4)	(a1-a3)
5	(c1,c2,c5)	(a1-a3-a4)
4	(c2,c3)	(a2-a4)

Table 2: Customer Demographic Data.

Tables 1 & 2 shows the different features of data that are involved in the debt clearance in the bank. Where data can be categorical, binary

2.4 Multi Method Combined Mining

The multi method mining is an approach to have the more informative knowledge in large data sources. This approach mainly focus on multiple data mining algorithms, which are needed to have efficient patterns that may be actionable rather than using single mining method in heterogeneous data sources.

In general there are multiple methods available to have patterns, these patterns can be found through the association rule mining method. Only by using of this method we cannot find actionable patterns to take decisions. In this paper we present two mining methods called association-classification methods serially to have actionable patterns.

In dealing with heterogeneous data the multi method mining involves in the following ways

- Firstly the domain knowledge, business understanding, data analysis and goal definition has to be taken into consideration.

- Secondly the patterns identified through mining methods.
- Finally after mining all patterns are combined and choose more workable patterns those suites to business environment.

The multi mining method work in the following three aspects:

2.4.1 Parallel multi method mining

2.4.2Serial multi method mining

2.4.3Closed loop multi mining method

2.4.1 Parallel multi method mining:

In this process the methods are parallel applied on given datasets to find atomic patterns and finally these patterns are combined.

2.4.2 Serial multi method mining: In this process the methods are applied in serial fashion that means one after the other to have more efficient patterns.

2.4.3 Closed loop multi mining method:

In this process either serial or parallel methods are applied iteratively and check whether the obtained patterns actionable or not, if not so the patterns are said back loop once again to have efficient patterns.

2.5 ID3 Method

This classification method takes frequent patterns and identifies the most actionable patterns from data sources through the measures of information gain. Using of these methods combined patterns are mined from Customer Demographic-Transactional data sources. The algorithm is as follows:

- ID3 (Examples, Target Attribute, Attributes)
- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root, with label = +.
- If all examples are negative, Return the single-node tree Root, with label = -.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Otherwise Begin ,A = The Attribute that best classifies examples. Decision Tree attribute for Root = A. For each possible value, v_i, of A, Add a new tree branch below Root, corresponding to the test A = v_i.
- Let Examples(v_i) be the subset of examples that have the valuev_i for A.

- If Examples(v_i) is empty ,Then below this new branch add a leaf node with label = most common target value in the examples.
- Else below this new branch add the sub tree ID3 (Examples(v_i), Target Attribute, Attributes – {A})
- End
- Return Root

The Proposed work can be applied to bank application in clearance of bank loan. Customers who had taken the loan were related to some other activities of bank, which cause sometimes fast clearance of loan or it may delay. Some activities are directly related to target and some may not. Hence these methods are used to find the actionable patterns that found in the distributed data sources. The following steps show the mining of combined patterns in the bank application.

1. According to domain knowledge and goal definition find the patterns *P* in data sources *S.*
2. Mine the necessary rule methods to mine the patterns.
3. Generate association rules to have frequent patterns *FP* by Apriori method.
4. These frequent patterns *FP* are passed serially through the Classification method called ID3 which generate the actionable patterns.
5. Finally all these rules are integrated to have efficient combined patterns CP.

2.6 Association-Classification Methods

A serial multi method mining method to have efficient patterns in the complex data.The Association-Classification rules are used to mine patterns from different heterogeneous data. Generally Association rule mining is used to generate patterns from data sources, these patterns may not feasible for decision-making. In addition to this we present another method called classification to have actionable knowledge.Association-Classification methods are applied in serial fashion. Apriori-ID3 algorithms to mine the combined patterns from heterogeneous data.

The first algorithm is used to mine frequent patterns and these patterns are serially applied to ID3 algorithm which identifies the more actionable patterns from complex data that makes business experts to have correct decisions.

2.6.1 Apriori method

Apriori is an association rule mining method is used to find frequent patterns from complex data. This method uses the support and confidence to measure the frequent patterns.

Pseudo-code for Apriori algorithm

- Join Step: Ck is generated by joining Lk-1with itself
- Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent k-item set

```
Ck: Candidate item set of size k
Lk: frequent item set of size k
L1= {frequent items};
for (k= 1; Lk!=∅; k++) do begin
    Ck+1= candidates generated from Lk;
    for each transaction tin database do
        Increment the count of all candidates in Ck+1that are contained in t
    Lk+1= candidates in Ck+1with min_support
end
return ∪kLk;
```

2.6.2 Methods to Improve Apriori Efficiency

The efficiency of the apriori algorithm can be improved by the following ways:

- Hash-based item set counting: A *k*-item set whose corresponding hashing bucket count is below the threshold cannot be frequent.
- Transaction reduction: A transaction that does not contain any frequent k-item set is useless in subsequent scans.
- Partitioning: Any item set that is potentially frequent in DB must be frequent in at least one of the partitions of DB.
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness.
- Dynamic item set counting: Add new candidate items sets only when all of their subsets are estimated to be frequent.

3.Conclusion

Enterprise applications deals with heterogeneous data sources, hence single step may not be feasible to mine the actionable patterns. The proposed concept of combined mining, for discovering the comprehensive data in complex data. Which focus on handling of data from multiple data sources, from multiple features or by multiple methods. The presented two mining methods called association-classification rules are used to have efficient mining of combined patterns from large heterogeneous data sources. The finally obtained combined patterns from these methods are efficient, which helps the business experts to take correct decisions in business environment so that they can enhance business profits and also the service objectives of the business.

4.Future Work

Future research will be developing effective paradigms, the interesting measures for handling large and multiple sources of data available in industry projects like government, stock market.

5.References

[1] Jeffrey W. Seifert-Data Mining: An Overview.

[2] Sotiris Kotsiantis, Dimitris Kanellopoulos- Association Rules Mining http://www.math.upatras.gr/esdlab/oldEsdlab/association%20rules%20kotsiantis.pdf

[3] Bing Liu Wynne Hsu Yiming Ma- Integrating Classification and Association RuleMining. http://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf

[4] Saso˘ Dzeroski- Multi-Relational Data Mining: An Introduction http://research.cs.wisc.edu/EDAM/Dzeroski.pdf

[5] Hong Cheng Xifeng Yan Jiawei Han Chih-Wei Hsu: Discriminative Frequent Pattern Analysis for effective classification.

[6] Langbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and chengqi Zhang "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE. Transactions on systems, man, and cybernetics,vol.41,no.3,june 2011.

[7]Y.Zhao,H.Zhang,F.Figueiredo,L.cao,C.Zhang,"Mining for Combined Association Rules on Multiple Datasets",inproc.DDDM,2007,pp.18-23

discovery,"IEEETrans.Know.Data.Eng.,vol.22,no.9,pp.1299-1312,sep.2010.

[8]L.Cao,Y.Zhao,D.Luo,andC.Zhang,"Flexible frameworks for actionable knowledge

[9]Y.Zhao,H.Zhang,L.Cao,C.Zhang,H.Bohlscheid,"Combined pattern mining:From learned rules to actionable knowledge,"in proc.AI,2008,pp.393-403.

[10]JieYin,Ling,C.X,Chen.T,"Post processing Decision Trees to Extract Actionable Knowledge,"pro.ACI,2003,pp685-688.