International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Effective of Combined Mining Techniques with Kinship Search in Complex Data

Narsheedha Beegum .P .A[1], Nimal J. Valath [2]

[1, 2] Calicut University, Vidya Academy of Science and Technology, Thalakottuara P.O, Thrissur, Kerala, India

Abstract: ***Enterprise data mining applications usually involve complex data such as multiple large heterogeneous data sources, users preferences, and business impact. A single method or one-step mining is limited in discovering informative knowledge in such situations. Mining complex data causes more space complexity and time complexity. It is crucial to develop effective approaches for mining patterns combining necessary information. The recent years have seen increasing efforts on mining more informative patterns. The project proposes combined mining as a general approach, rather than presenting a specific algorithm. Combined mining is a framework for mining informative patterns by combining components from either multiple data sets or multiple features or by multiple methods on demand. Proposed general frameworks, paradigms, and basic processes for multi-feature combined mining, multi-source combined mining, and multi-method combined mining using existing algorithms. The project implements combined mining approach with kinship search technique to generate the effective patterns in medical field. By observing informative patterns acquired from the above techniques efficient actionable decision making is possible.***

Keywords: Combined mining, Complex data, Data mining, Pattern Mining, Data set, Item set

1. Introduction

In data mining in order to discover knowledge, general framework is used, called as knowledge discovery in database (KDD). Generally use extraction of association rules for data mining. Now researchers focus on extracting informative knowledge in complex data. A pattern is considered as informative if user can act upon it for his advantage. Real time complex data consists of vast information. For mining effective patterns, existing single traditional data mining method is not enough.

It proposes 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 medical field data sets. Flexible frameworks for combining multi features, multi sources, and multi methods covering various needs in mining complex data, which are customizable for specific cases. In the medical and pharmaceutical area it is important to know whether a patient will react positively or negatively to a treatment or a drug. Different treatments can be performed on different patients based on their diseases. So the main objective of the project is to implement combined mining approach in medical field data which helps the researchers to find which are the best suggested medicines for different diseases based on the recovery of patient, combined mining approach implemented with different data mining algorithms on medical field data set.

Until now, rather simple statistics have mostly been used in medicine/pharmacy for such problems. Data Mining offers the potential for much deeper analysis and predictions in this field. Any medical attributes are non-numeric which further makes Data Mining a better choice in comparison to traditional statistics tools. This helps the medical and pharmaceutical industry, but first and foremost of course the patients. Patterns identified by traditional methods usually only involve homogeneous features from a single source of data, e.g., frequent patterns of customer shopping habits. Such patterns consist of a single line of information and are not informative in business decision making. If attributes from multiple aspects can be included, the resulting patterns can then completely reflect the business situation and be workable in supporting business decision making.

2. Existing Work

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. First, most of existing single-handed data mining methods do not target the discovery of informative patterns in complex data, as discussed in this paper. 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.

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 due to the time and space limit. In addition, techniques for involving multiple methods and handling multiple data sources are often specifically developed for particular cases. A typical challenge is that a huge amount of sequential patterns is usually mined in the sequential mining procedure. Although pruning algorithms are used for post processing, there is still a large amount of sequential patterns constructing the feature space. Moreover, existing algorithms often do not tackle important problems such as how to efficiently and effectively select

discriminative features from the large feature space. This issue is handled in our closed-loop sequence classification method.

2.1 Recent Approaches for pattern mining and Association rule mining

Wei Fan, Kun Zhang, Hong Cheng, Jing Gao, Xifeng Yan, jiawei Han, Philip S. Yu, and Olivier Verscheure [1] proposed a frequent pattern mining in two steps. Paper proposes a new different method to find highly compact and discriminative patterns. It builds a decision tree having different nodes and each node identify discriminative patterns with minimum support. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu proposed an approach to mine sequential pattern-growth approach for efficient mining of sequential patterns [2]. By this approach sequence database is recursively projected into a set of smaller projected databases, and sequential patterns are grown in each projected database by exploring only locally frequent fragments. Based on the pattern growth sequential pattern mining paper proposed a most efficient method pefixspan which offers reduced projected databases. Pseudoprojection is also used in prefixSpan. Yanmin Sun, YangWang, and Andrew K.C.Wong, Fellow proposed a method for Association mining and classification [3]. Paper proposed AdaBoost algorithm indicates that boosting simple rules could often achieve better classification results than the use of complex rules.

Longbing Cao, Yanchang Zhao and Chengqi Zhang proposed Algorithms which are developed to mine frequent positive-impact-oriented and negative impact- oriented activity patterns, sequential impact-contrasted activity patterns and sequential impact reversed activity patterns [4]. K. K. Rohitha G. K. Hewawasam, Kamal Premaratne, and Mei-Ling Shyu proposed an association rule mining based classification algorithm on Dempster Shafer belief-theoretic relational Database[5]. Various ARM-related notions are revisited so that they could be applied in the presence of data imperfections. Longbing Cao, Yanchang Zhao, Huaifeng Zhang, Dan Luo, Chengqi Zhang and E.K. Park proposed a process AKD (actionable knowledge discovery) which itself is a closed optimization problem solving process from problem definition, framework/model design to actionable pattern discovery, and is designed to deliver operable business rules that can be seamlessly associated or integrated with business processes and systems [6].

Mr Rajesh K Ahir and M.S Mithal.B Ahir in paper Algorithm for mining frequent patterns: a comparative study proposed a concept of comparing the frequent patterns obtained by using two different efficient algorithms, apriori algorithm and FP growth algorithm. Comparing the result obtained by the two algorithms and result is displayed with the help of bar charts [7].

3. Combined Mining

Rather than presenting a specific algorithm for mining a particular type of combined patterns, work focuses on abstracting several general and flexible frameworks from the architecture 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. Combined mining is classified as shown in figure 1.

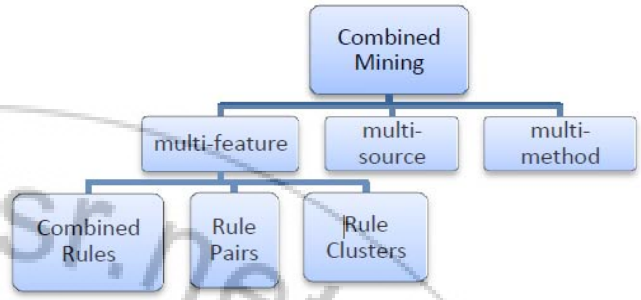

Figure 1: Classification of Combined mining

The general ideas of combined mining are as follows.

a) By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.
b) By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.
c) Effective of Combined Mining Techniques with Kinship Search in Complex Data
d) 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.
e) By applying multiple interestingness metrics in pattern mining, patterns are generated which reflect concerns and significance from multiple perspectives

3.1 Modular Description

There are four modules included
A. Multi-source Combined Mining
B. .Multi-feature combined mining
C. Multi-method combined mining
D. Kinship search

A. Multisource Combined Mining

Mining complex data requires the handling of multidata sources implicitly or explicitly. It supports the discovery of combined patterns either in multiple data sets or subsets (D1, . . . ,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 D1) are selected for mining exploration (R1); 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 deployment of method Rk (k = 2,. . . , L), which could be either in parallel or through combination, is informed by 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 (PRn) identified from individual data sets are merged (GPn) with the involvement of domain knowledge and further extracted into final deliverables (P).

B. Multifeature Combined Mining

In multifeature 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. Based on the different expectations on combined pattern types, MFCPs may be instantiated into pairs, clusters, incremental pairs, and incremental clusters. Correspondingly, the discovery of such types of patterns can be segmented into six steps on demand. The process is as follows. First, atomic patterns P1 are discovered in one data set and then are used to partition another data set. Then, in a derived subdata set, atomic patterns P2 are discovered. After that, P1 and P2 are merged into a combined pattern. Through finding common prefixes or postfixes in these patterns, interesting pair patterns are discovered by putting contrast patterns together. In addition, patterns with the same prefixes or postfixes form cluster patterns. Finally, incremental pair and cluster patterns can be built upon the identified pairs/clusters, respectively. This phase handles the ontology concepts of each web service. The web service file which was selected in the previous phase that files corresponding ontology file is displayed here. For each ontology file, its concepts, relationships, properties and keyword count is evaluated.

C. Multimethod Combined Mining

The general process of multimethod combined mining is as follows.

1) First, based on the domain knowledge, business understanding, data analysis, and goal definition, a user determines which methods should be used in the framework.
2) Second, the patterns discovered by each method are combined with the patterns by the other methods in terms of merging method G. In reality, the merger could be through either serial or parallel combined mining.
3) Finally, after mining by all methods, the combined patterns are further reshaped into more workable patterns.

a) Parallel Multimethod Combined Mining: One approach to involving multiple methods for combined mining is the parallel multimethod combined mining. In parallel multimethod combined mining, multiple methods are implemented on multiple data sources or partitioned data sets. The resulting patterns are the combination of the outputs of individual methods on particular data sources.
b) Serial multimethod Combined Mining: The second type of approach to involving multiple methods into combined mining is the serial multi method combined mining, which is described as follows. In serial multimethod combined mining, the data mining methods are used one by one according to specific arrangements. That is, a method is selected and used based on the output of the previous methods. Such serial combination of data mining methods is often very useful for mining complex data sets.
c) Closed-Loop Multi method Combined Mining : In serial multimethod combined mining, a previously applied method Rj , in general, has no impact on another methods (Ri) resulting patterns and performance, even though Rjfollows Ri. This is actually a common issue in openloop combination. In practice, the feedback from latter methods results to its previous methods may assist with the pattern refinement in combination and enhance the deliverable performance and the efficiency of the data mining process. To this end, we propose the concept of closed-loop multimethod combined mining.
d) Closed-Loop Sequence Classification: In order to build sequential classifiers, a number of processes, such as the significance test and the coverage test, have to be conducted on the sequential pattern set. If the sequential pattern set contains huge amounts of sequential patterns, the classifier building can also be extremely time consuming. Therefore, in sequence classification, the efficiency problem exists not only in sequential pattern mining but also in classifier building

Apriori algorithm and fp tree algorithm is used for ultimethod combined mining. Apriori Algorithm operates on a list of transactions containing items (different measures like disease, age, blood group etc on different tables in hospital data). Frequent occurrences of items with each other are mined by Apriori to discover relationship among different items. A single transaction is called an Item set. Apriori uses a minimum support value as the main constraint to determine whether a set of items is frequent. In the first pass of the algorithm, it constructs the candidate 1-itemsets.The algorithm then generates the frequent 1-itemsets by pruning some candidate1-itemsets if their support values are lower than the minimum support. After the algorithm finds all the frequent 1-itemsets, it joins the frequent 1-itemsets with each other to construct the candidate 2-itemsets and prune some infrequent item sets from the candidate 2-itemsets to create the frequent 2-itemsets. This process sis repeated until no more candidate item sets can be created. We are calculating other measure called interestingness for each pattern by calculating support, lift and confidence by the algorithm. In selecting cluster type of patterns traditional contributions, support, lift and confidence is limited. By calculating interestingness values of each atomic, pair patterns within the cluster pattern, interestingness value of entire cluster pattern is calculated. So after apriori algorithm we are obtaining frequent cluster patterns with its interestingness value

FP tree algorithm: The frequent-pattern tree is a compact structure that stores quantitative information about frequent patterns in a database. Method is as follows. 1) Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items. 2) Create the root of an FP-tree, T, and label it as “null”. Then do first, select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Secondly Call insert tree ([p | P], T). The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name,

then increment N 's count by 1; else create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N) recursively.

D. Kinship Search

A kinship search will test the relationship between two or more individuals to assess if they are biologically related. In this search, we can search the query as the blood relations query. Our proposed system provides valuable or effective result for this type of query according to the results generated from the combined mining

4. Dataset Description

The data set considered here is a medical database to find out quality rules in order to make decisions to find which are the best suggested medicines for different diseases. The data set consists of following attributes.

Table 1: Dsease Table

Attributes	*Description*
Report_id	Patent d
Disease	Name of the dsease
Symptoms	One or more symptom names

Table 2: Medcne Table

Attributes	*Descirption*
Medcne	Medicine name
Disease	Name of the disease
Class	Class which medicine comes

Table 3: MedicineForm Table

Attrbutes	*Descirption*
Medicine	Medicine name
Dosage Form	Dosage of medicine
Medicine involved	Content of medicine involved

Table 4: PatientStatus Table

Attributes	*Description*
Report_id	Patient id
Gender	Female or male
Seriousness	Seriousness of disease (yes/no)
Recovery_status	Recovered/resolved
Death	death of patient or not

Table 5: Patient Table

Attributes	*Descirption*
Report_id	Patent id
Gender	Female or male
Age	Age in number
Seriousness	Seriousness of disease (yes/no)

5. Experimentation Performed

The below figures(234) shows which are the best suggested medicines with recovery values using multisource , multi-feature and multi-method combined mining approach. Higher the value higher the recovery rate for that particular medicine. The datasets are taken as the input. Combined patterns are generated using multifeature multisource and multimethod combined mining. Search option is given in the text box with disease name and symptoms. The results are generated in graph which shows which are the suggested medicines for that disease with the given symptoms. The output is taken from the combined patterns generated using the three approaches of combined mining.

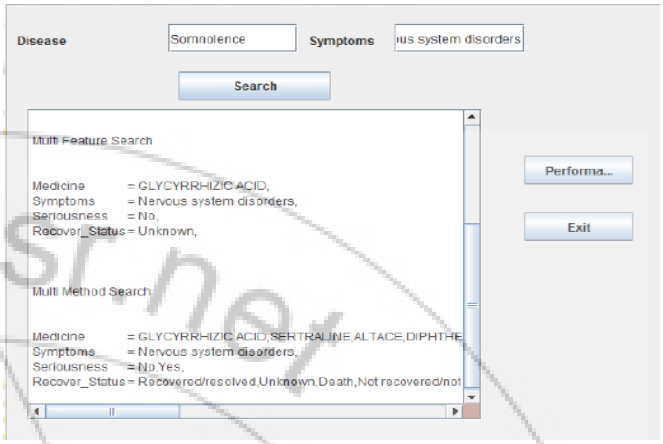

Figure 2: kinship search

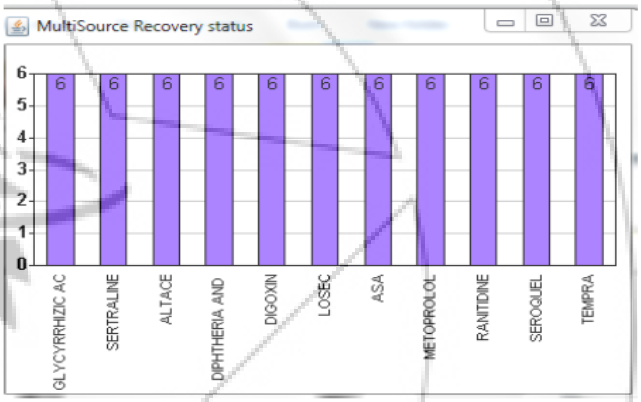

Figure 3: Multi-source Combined Mining

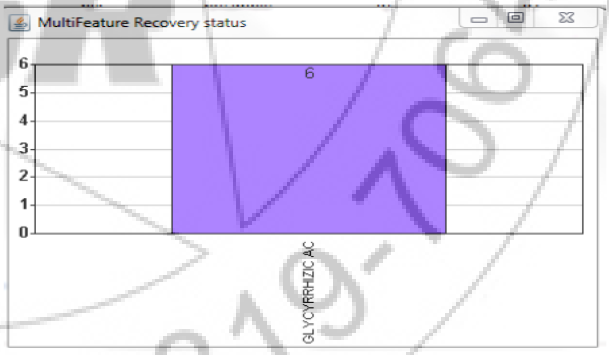

Figure 4: Multifeature Combined Mining

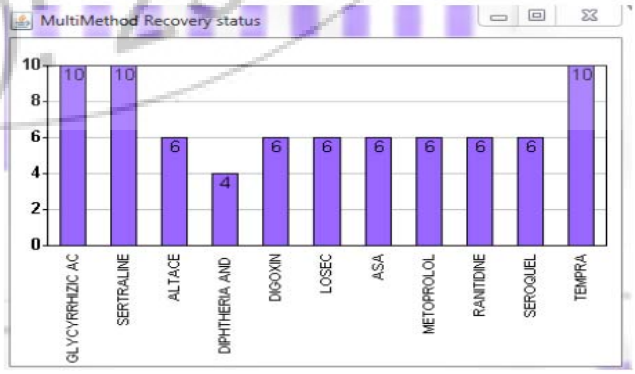

Figure 5: Multi-method combined mining

Comparing the performance of the proposed system such as multisource combined mining with kinship search, multifeature combined mining with kinship search and multimethod combined mining with kinship search in terms of time complexity. Multimethod Combined Mining is more efficient when comparing the other two mining approaches.

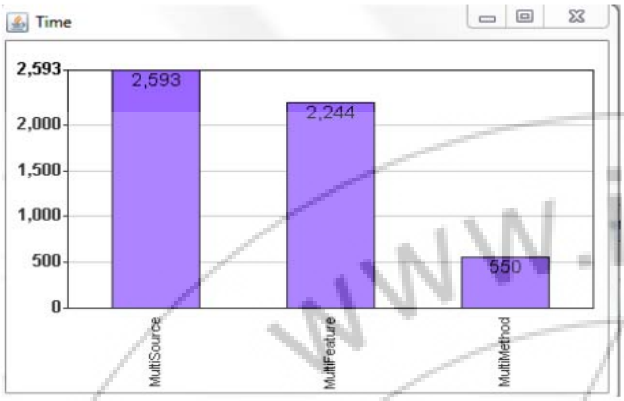

Figure 6: Performance graph of three combined mining techniques

6. Conclusion

Paper has presented a comprehensive and general approach named combined mining for discovering informative knowledge in medical data. Work mainly focuses on discussing the frameworks for handling combined mining, such as multifeature, multisource, and multimethod combined mining. The system addressed challenging problems in combined mining such as space complexity, time complexity, limitation of source and heterogeneous nature of dataset .The system also proposed effective pattern merging and interaction paradigms. Generated combined pattern types, such as pair patterns and cluster patterns with interestingness measures. Used an effective tool - dynamic chart for presenting complex patterns in a business-friendly manner. In this proposed system, introduced the kinship search technique. A kinship search will test the relationship between two or more individuals to assess an alleged relationship. Comparison of the performance of multisource, multifeature and multimethod combined mining is also done by using time complexity.

References

[1] Wei Fan, Kun Zhang, Hong Cheng, Jing Gao, Xifeng Yan, Jiawei Han, Philip S. Yu, and Olivier Verscheure, ”Direct Mining of Discriminative and Essential Frequent Patterns via Model-based Search Tree, IEEE Transactions on Systems, Man and Cyberneticspart B: CYBERNETICS, vol. 41, no. 3, june 2011, pp. 699-71R.

[2] 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 Transactions Knowledge Data Eng,2004,pp. 1424-1440.

[3] Y. Sun, Y.Wang, and A.Wong, Boosting an associative classifier, IEEE, 2006,pp. 988-992.

[4] L. Cao, Y. Zhao, and C. Zhang,Mining impact-targeted activity patterns in imbalanced data, IEEE Trans. Knowl. Data Eng., vol. 20, no. 8, 2008, pp. 1053-1066..

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

[6]] 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. 9,2010,pp. 1299-1312.

[7] Mr Rajesh K Ahir and M.S Mithal.B Ahir , ”Algorithms for mining Frequent Patterns: A comparative Study”,International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12,2013.

[8] Longbing 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 part B: CYBERNETICS, vol. 41, no. 3,2011, pp. 699-712.

[9] Lingjuan Li, Min Zhang , The Strategy of Mining Association Rule Based on Cloud Computing ,International Conference on Business Computing and Global Informatization , 2011,pp.475-478

[10] Goulbourne, G., Coenen, F. and Leng, P. (2000), Algorithms for Computing Association Rules Using a Partial-Support Tree, Journal of Knowledge- Based Systems, Vol (13), pp141-149.

[11] Coenen, F., Goulbourne, G. and Leng, P., (2003). Tree Structures for Mining association Rules, Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp25-51.

Author Profile

Narsheedha Beegum received the B.Tech from Calicut University in 2008 and Continuing M.Tech degrees in Computer Science and Engineering from Vidya Academy of Science and Technology in 2014, respectively. This paper is a part of research done in combined data mining. Her research interests include data mining and machine learning and their applications

Nimal J Valath received the B.Tech. from Cochin University in 2011 and Continuing M.Tech degrees in Computer Science and Engineering from Vidya Academy of Science and Technology in 2014, respectively. This paper is a part of research done in combined data mining. His research interests include combined pattern mining, sequence analysis, behavior analysis and natural language processing.