

Effective of Combined Mining Techniques with Kinship Search in Complex Data

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

¹M.Tech Student,²M.Tech Student

¹Department of Computer Science,
Vidya Academy of Science and Technology, Kerala, India

Abstract—Enterprise data mining applications often involve complex data such as multiple large heterogeneous data sources, users preferences, and business impact. In such situations, a single method or one-step mining is often limited in discovering informative knowledge. 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, e.g., integrating frequent pattern mining with classifications to generate frequent pattern-based classifiers. 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, multisource combined mining, and multimethod combined mining. 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.

Index Terms—Combined Mining, Pattern Mining, Data set, Item set

I. 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. To acquire knowledgeable information from data source we should integrate one or more data mining methods. Combined mining approach implemented with different data mining algorithms on medical field data set.

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 multifeatures, multisources, and multimethods 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.

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.

II. EXISTING SYSTEM

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) postanalysis and postmining 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 postprocessing, 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

III. PROPOSED METHOD

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 5
- 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.

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. Proposed system helps the researchers to find which are the best suggested medicines for particular disease by analyzing the result from medical field dataset. It is effective and provide more interesting and actionable information than traditional association rules. To improve the efficiency data could be distributed across multiple datasets and multiple data sources and then the computations can be carried out in parallel. Kinship technique is applied to get efficient informative patterns of data. These informative patterns are used as per for user preference to take actionable decision. A kinship search will test the relationship between two or more individuals to assess if they are biologically related.

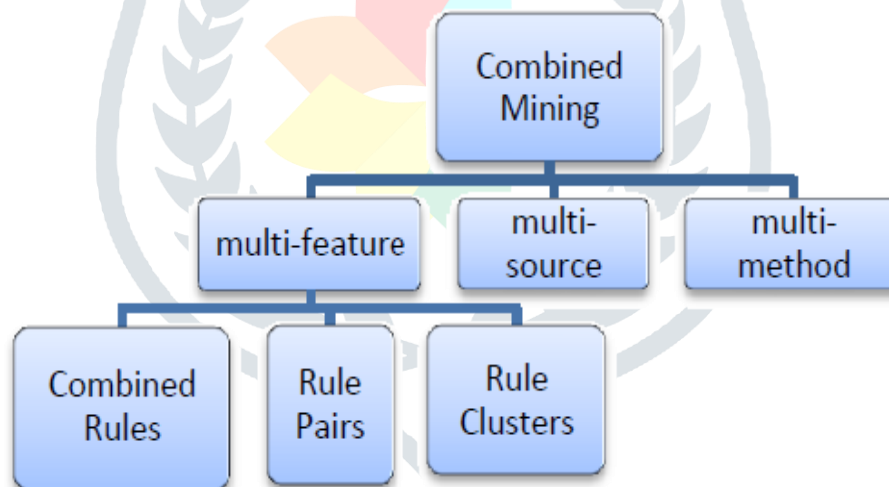


Fig-1 Classification Of Combined Mining Approach

3.1 MODULAR DESCRIPTION

- A. Multisource Combined Mining
- B. Multifeature combined mining
- C. Multimethod combined mining
- D. Kinship search

A. Multisource Combined Mining

Mining complex data requires the handling of multiple data sources implicitly or explicitly. It supports the discovery of combined patterns either in multiple data sets or subsets (D_1, \dots, D_K) 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 D_1) are selected for mining exploration (R_1); 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 R_k ($k = 2, \dots, L$), 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 D_k with the supervision of the results from step $k-1$; and 3) after finishing the mining of all data sets, patterns (PR_n) identified from individual data sets are merged (G_{P_n}) 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 P_1 are discovered in one data set and then are used to partition another data set. Then, in a derived subdata set, atomic patterns P_2 are discovered. After that, P_1 and P_2 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.

- **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.
- **Serial multimethod Combined Mining:** The second type of approach to involving multiple methods into combined mining is the serial multimethod 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.
- **Closed-Loop Multimethod Combined Mining :** In serial multimethod combined mining, a previously applied method R_j , in general, has no impact on another methods (R_i) resulting patterns and performance, even though R_j follows R_i . This is actually a common issue in open loop 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.
- **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 multi method combined mining

- **Apriori Algorithm:** It 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 Itemset. 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 candidate 1-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 itemsets from the candidate 2-itemsets to create the frequent 2-itemsets. This process is repeated until no more candidate itemsets 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 contribution, 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.

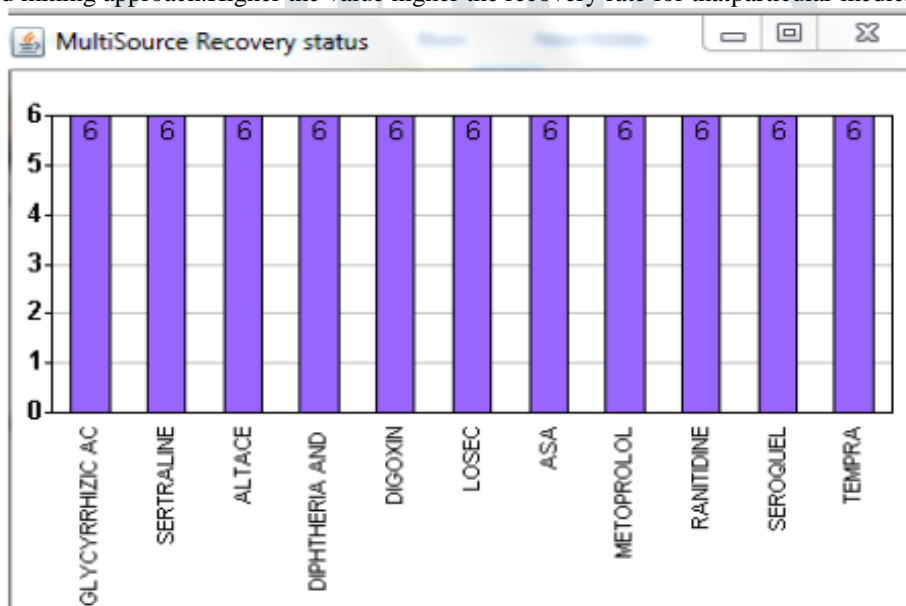
D Kinship Search

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.

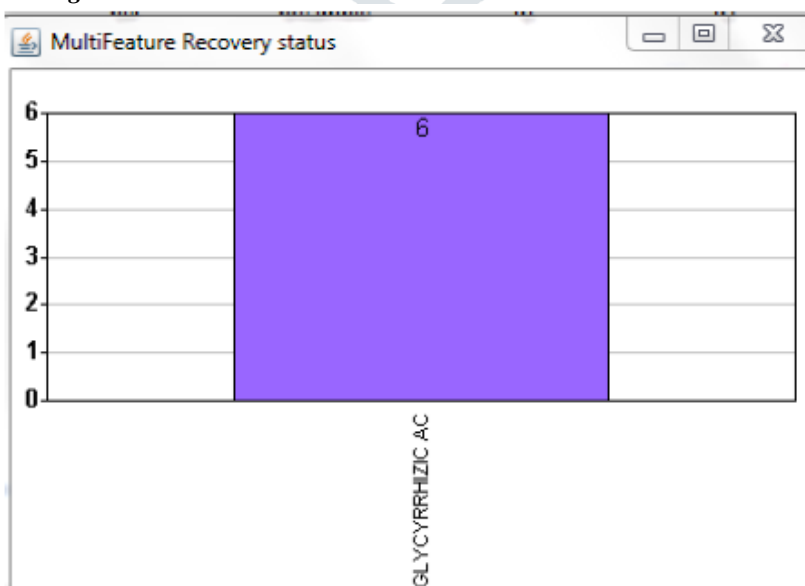
IV. SIMULATION AND RESULT

A. Multisource Combined Mining

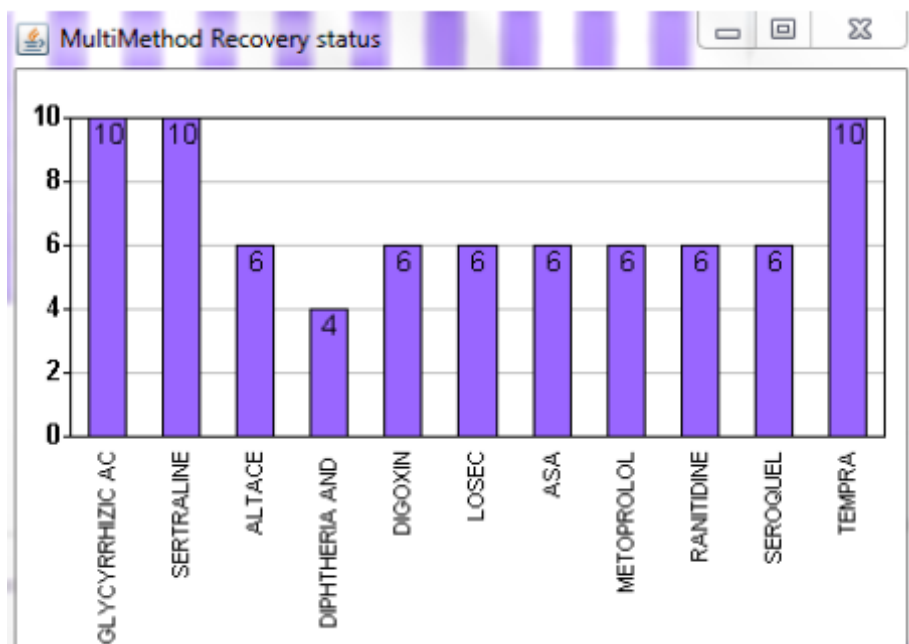
The below graphs shows which are the best suggested medicines with recovery values using multisource, multi-feature and multi-method combined mining approach. Higher the value higher the recovery rate for that particular medicine.



B. Multifeature Combined Mining

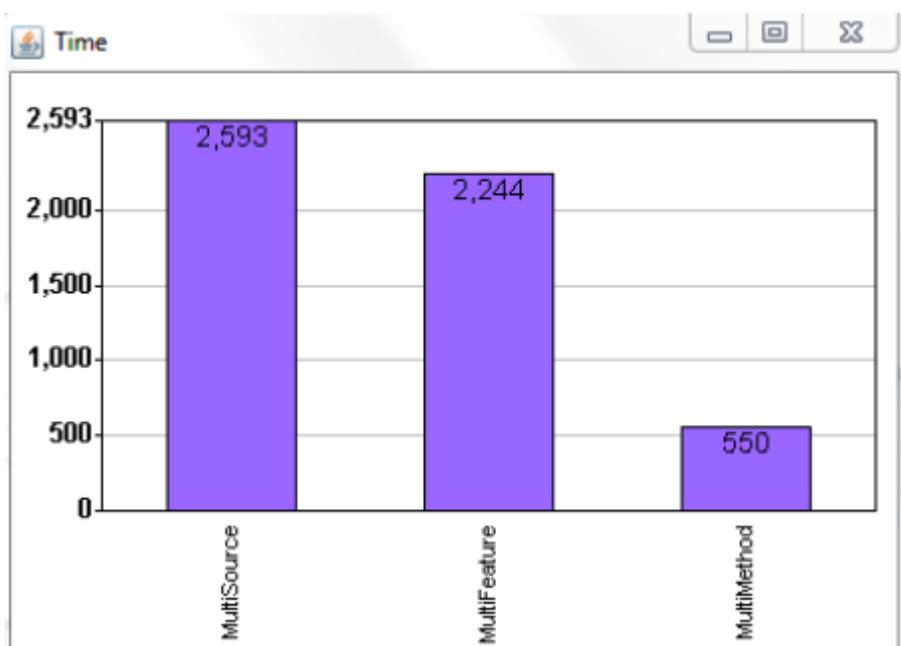


C. Multimethod Combined Mining



D. Performance of Three 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.



V.CONCLUSION

Project 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.

VI. 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 Cybernetics part B: CYBERNETICS, vol. 41, no. 3, June 2011, pp. 699-71.
- [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] 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, 2007, pp. 1446-1459.
- [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. BAhir, "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), pp. 141-149.
- [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, pp. 25-51.

