ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Combined Mining: An Approach for Actionable Pattern Discovery in Complex Data

Seema A. Shabadi[1]
PG Student
Department of Computer Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, Maharashtra – India

Dr. A. B. Bagwan[2]
Professor and HOD
Department of Computer Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, Maharashtra – India

Abstract: ***Data mining for applications of big business enterprise often involve complex data, having heterogeneous data sources, user preferences & different business needs. A single step or one method mining, often become limited in discovering informative patterns in such situations. Joining relevant large data sources for pattern mining which consists of different aspects of information turn into time & space consuming process. It is strongly needed to develop effective ways for mining patterns by combining necessary information from multiple relevant business lines, gathering for real business settings & decision making actions rather than just providing a single line of patterns. The frameworks can be built by combining components from either multiple data sets or multiple features or by multiple methods. This paper proposes the approach to efficient mining in the form of combined mining methods, combined pattern types, pattern merging methods & interestingness measures which are effective paradigms for handling large & multiple sources of data available in various sectors like government, stock market, insurance & banking. In combined mining approach Probabilistic Lossy-counting algorithm is used on each data-source to get the frequent data item-sets and then get the combined association rules. In multi-method combined approach, incremental mining algorithm to enable Probabilistic Frequent Item set (PFI) and Frequent Pattern (FP) Tree algorithm are combined to make a classifier to get more informative and actionable pattern. In multi-feature combined mining approach, pair patterns and cluster patterns are obtained to generate incremental pair patterns and incremental cluster patterns, which cannot be generated directly by the existing methods.***

Keywords: ***Association Rule Mining, Combined Mining, Complex Data, FP-tree, Incremental Cluster-Patterns, Incremental Pair-Patterns, Multiple Source Data Mining, Probabilistic Frequent Item-set.***

I. INTRODUCTION

Complex and large scale data sets, such as public service data often involves multiple, distributed, and heterogeneous data sources, which contains the information about business transactions, user preferences, and business impact. In such situations, business people indeed expect the discovered knowledge to present overall picture of business outline rather than one view based on single source. Data sampling generally not accepted because it may miss some important data that may be filtered out during sampling. Joining of tables may not be possible due to the time and space limit. More often this approach of handling multiple data sources can only be developed for specific cases and cannot be applied for all problems.

Combined mining is a two-to-multistep data mining approach, which involves first mining the atomic patterns from each individual data source and then combines those atomic patterns into combined-patterns by pattern-merging method, which is more suitable for a particular problem. In multi-source combined mining approach, the informative patterns from individual data source are found and then the combined patterns are generated, which can't be directly generated by some traditional algorithms like FP-growth etc. In multi-feature combined mining approach, features from multiple data sets are considered while generating the informative patterns, where it is necessary in order to make the patterns more actionable. In case of cluster patterns, the

cluster of patterns with same prefix is made but the remaining data items in the pattern make the results to be different. The main advantage of this approach is neither any pruning method is applied nor any clustering method separately. To get the more informative patterns during the probabilistic lossy-counting algorithm's implementation, pruning is done at the boundary of the data sources and the most similar data items in the same bucket itself are extracted. The combined mining approach is used to directly identify patterns enclosing elements from multiple sources or with heterogeneous features. Its deliverables are combined informative and actionable patterns which consist of multiple components, a cluster or pair of atomic patterns, found in individual sources or based on individual methods.

In this paper we have focused on obtaining the patterns with FP Tree which then combined with PFI mining to get the effective patterns. Subsequently these patterns are used for incremental mining which will consider the patterns lying on boundary for which probability is near to the threshold value, which cannot be obtained by only single mining like FP Tree. A combined association rule is composed of multiple heterogeneous item sets from different data sets. The work shows that such combined rules cannot be directly produced by traditional algorithms.

II. LITERATURE SURVEY

L. Cao, H. Zang, Y. Zhao, Dan Luo, C. Zhang [1], this paper addresses comprehensive and general approach named combined mining for discovering informative knowledge in complex data. They have focused on discussing the frameworks for handling multifeature-,multisource-, and multimethod-related issues. They have addressed challenging problems in combined mining and summarized and proposed effective pattern merging and interaction paradigms, combined pattern types, such as pair patterns and cluster patterns, interestingness measures, and a dynamic chart for presenting complex patterns in a business-friendly manner. The frameworks are extracted from their relevant business projects conducted and currently under investigation from the domains of government service, banking, insurance, and capital markets. 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.

C. Zhang et al [3], proposed a novel approach of combined patterns to extract important, actionable and impact oriented information from a large amount of association rules. They also proposed definitions of combined patterns and also design novel matrices to measure their interestingness and analyzed the redundancy in combined patterns.

Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang [4], this paper shows a framework for mining combined association rules from multiple datasets, with heterogeneous datasets requiring different data mining techniques capable of producing comprehensive and useful rules. This framework has been tested with real-world social security data and the results are interesting and help business to classify customers as quick/moderate/slow payers and their repayment patterns.

H. Zhang, Y. Zhao, L. Cao, and C. Zhang [7], this paper proposes an efficient algorithm to mine combined association rules on imbalanced datasets. Unlike conventional association rules, our combined association rules are organized as a number of rule sets. In each rule set, single combined association rules consist of different kinds of attributes. A novel frequent pattern generation algorithm is proposed to discover the complex inter-rule and intra-rule relationships. Data imbalance problem is also tackled in this paper. The proposed algorithm is tested in a real world application.

Long bing Cao [14], This paper presents a high level picture of combined mining, and discusses many novel aspects of pattern relation analysis and combined patterns. Pattern combination dimensions, pattern combination criteria, pattern relations, pattern structures and pattern paradigms, which are important for constructing combined patterns and for discovering actionable knowledge in complex data, are discussed. Pattern ontology and the pattern dynamic chart are also introduced to present

combined patterns. Combined pattern and combined mining present a general paradigm with great potential for identifying and producing informative and actionable patterns.

B. Liu, W. Hsu, and Y. Ma [16], proposed a technique which first prunes the discovered association-rules to remove the insignificant association-rules from the entire set of association- rules, and then finds a subset of the un-pruned association-rules by which a summary of the discovered association-rules can be formed. They called that subset of association-rules as the direction setting rules because they can be used to set the directions, which are followed by the rest of the association-rules. By the help of the summary, the user can have more focus on the important aspects of the particular domain and also can view the relevant details. They suggest that their approach is effective as their experimental result shows that the set of direction setting rules is quite very small.

Kargupta et al [17], presented a framework of collective data mining to conduct distributed data mining from heterogeneous sites. They point out that in a heterogeneous environment, naïve approaches to distributed data analysis may lead to incorrect data-model.

Karypis and Wang [18], presented a new classifier, HARMONY, which is an example of direct mining for informative patterns as HARMONY directly mines the resultant set of rules required for classification.

III. Implementation Details

A. *Problem Definition*

Complex data may contain implausible information, which cannot be mined directly just by using a single method, and also it is tough to deal with such information using different perspective such as client's perspective, business analyst's perspective and decision-makers perspective, etc. due to increased level of complexity. A framework is to be developed to mine for comprehensive & informative patterns in such complex data from multiple data sources with combined pattern types, combined mining methods, pattern merging methods & interestingness measures with effective paradigms for handling large and multiple data sources. The approach used here is, to try to get patterns to retrieve useful information and patterns from complex data. This information can be used in different places, for example in government service related data, data from social security, e-commerce, stock market, market campaigns, measuring the success of marketing efforts and client-company behaviour, etc.

B. *Block Diagram*

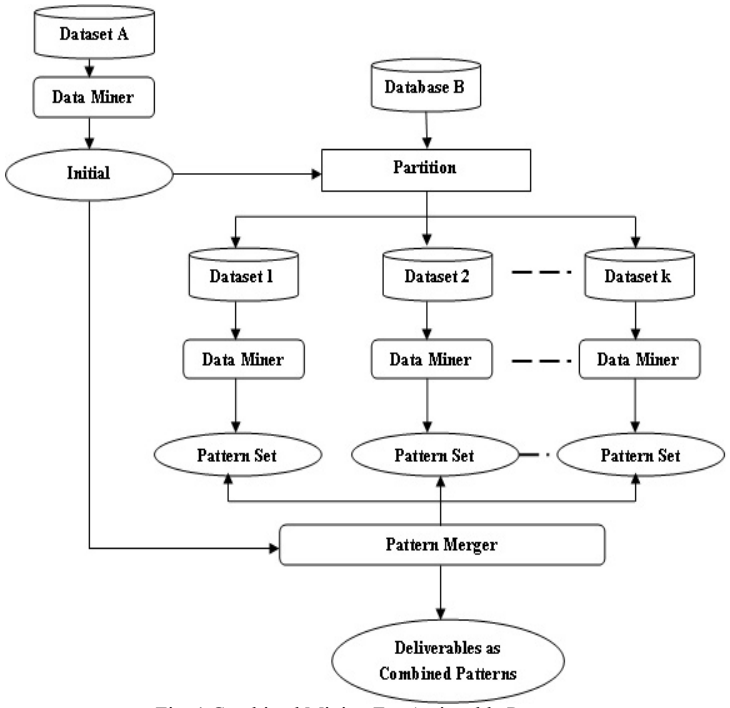

Fig. 1 Combined Mining For Actionable Patterns

C. Proposed Solution

The type of data used during the pattern discovery process determines the quality and effectiveness of the patterns. Our proposed solution, according to the problem definition consists of following steps:

Preprocessing and Data Mining Approach: First remove the tuples which contain an unknown value for any of the attribute in a dataset, because these kind of tuples are source of noise or errors. Second, make non-overlapping partitions of dataset to form sub datasets. The reason behind generating sub-datasets is that, each of the sub-dataset can be used as source of data for multi-source combined mining approach. The information needed for the classification of a tuple in a partition *P*, if the class label attribute has *n* distinct values, each of which defines a distinct class [2], as follows:

Where, p_i is the probability that a particular tuple in partition *P* belongs to class C_i and we can compute p_i as $p_i = |C_{i,p}|/|P|$ and where $C_{i,p}$ defines the set of tuples of class C_i in *P* while $|C_{i,p}|$ and $|P|$ denotes the number of tuples in $C_{i,p}$ and *P* respectively.

$$I(P) = -\sum p_i \log_2 p_i \qquad \text{Where } i = 1..n \qquad (1)$$

Then tuples in *P* are classified on some attribute *A* having m distinct values, which we get from training dataset. Then the amount of information needed for an exact classification [2] is measured by

$$I_A(P) = \sum(|P_j|/|P|) * I(P_j) \text{ Where } j = 1..m \qquad (2)$$

Then the information *G(A)* for attribute *A* is measured as:

$$G(A) = I(P) - I_A(P) \qquad (3)$$

For computing the information gain for continuous attributes *B* in a partition *P*, we have to compute the information gain for every possible split-point for *B* and then choosing the best split- point. We consider split-point as a threshold on *B* First, we have to sort the values of *B* in increasing order and then typically the mid-point between each pair of adjacent values considered as a possible split-point. So, if there are *y* values of *B*, then (*y-1*) possible splits has to be computed. The reason for sorting the values of *B* is that if the values are already sorted then for determining the best split for *B* requires only one pass through the values. Then, for each possible split-point for *B*, by using equation (2), we measured $I_A(P)$, where the number of value of m = 2. The point with the minimum expected information requirement for *B* will be selected as the best split-point for *B*.

After finding the information gain for each attribute, we have taken an attribute with maximum information gain from each partition P_i and then by concatenating the attribute values from all partitions.

Second, remove the duplicate records from the data-set, which will increase the efficiency of mining. We form a data-stream, which serves as an input for probabilistic lossy-counting algorithm.

Probabilistic Lossy Counting Algorithm (PLC): The error bound associated with an element present in a table determines which elements to remove from the table. An element is removed from the table if the sum of its frequency and error bound is less than or equal to a given threshold. This causes elements with a large error bound remaining in the table over many windows. The error bound value has a straight impact on the memory consumption of the algorithm. PLC [12], provides probabilistic guarantees and is useful to make the error bound significantly smaller than deterministic error bound of lossy counting. This will result in having a table for fewer windows and less memory consumption. From the studies it has been observed that a probabilistic error bound can be substantially smaller than a deterministic error bound.

PLC computes frequency counts over a stream of data-items. Frequency of item-sets is approximated within a user-specified error bounds E_{rr}. If N is the current length of the data-stream then this algorithm takes $1/ E_{rr}\log(E_{rr}N)$ space in worst-case for computing the frequency counts of a single data-item. Following are the implementation steps of this algorithm:

Input: Support s, error bound E_{rr} and input data-stream

Output: Set of data-items with frequency count at least equals to

$$(s- E_{rr})N$$

Step 1: Divide input data-stream logically into the buckets of width w=ceil(1/ E_{rr}) and each bucket is labeled with bucket id, initially starting from 1 for the first bucket, and the current bucket id is denoted by $B_{Current}$, which equals to ceil(N/ E_{rr})

Step 2: Maintain a data structure DS, which is a set of values of the form (E, FE, δ), where E is an element from the input data-stream and FE is the true frequency of the element E and δ denotes the maximum number of times E could have occurred in first $B_{Current}$ -1 buckets. Initially DS will be empty.

Step 3: For an element from data-stream, if E already exist in DS then increase its FE by 1 else we have to create a new entry in DS such as (E, 1,BCurrent -1).

Step 4: If it is the bucket boundary then we have to prune DS as follows:

$F_E + \delta <=$ BCurrent then the entry from DS which is (E,F_E,δ) has to be deleted

Step 5: When a user wants a final list of frequent data-items with support s, then output all those entries in DS with FE (s- E_{rr})N.

Then, we generate the combined association rules (C. Zhang et al, 2011) of the frequent data- items computed by probabilistic lossy-counting algorithm.

Multi-feature Mining Approach: After generating the combined association rules, heterogeneous features of different data types as well as of different data categories are considered. If the combined association rule is of the form IF *R* THEN *S*, where *R* is the antecedent and S is the consequent part of the rule, then we have some traditional definitions for support, confidence and lift of the rule as given below in the Table I.

TABLE I Support, Confidence and Lift for the rule R→S

SUPPORT	Prob(R U S)
CONFIDENCE	Prob(R U S)/Prob(R)
LIFT	Prob(R U S)/(Prob(R)*Prob(Y))

On the basis of these traditional definitions of support, confidence and lift, we can compute the Contribution and Interestingness [1] I_{RULE} of the rule R_X U R_Y → S as follows:

Contribution (R_X U R_Y → S) = LIFT (R_X U R_Y → S)/ LIFT (R_X → S)

$$=\text{CONFIDENCE}(R_X \text{ U } R_Y \rightarrow S)/\text{CONFIDENCE } (R_X \rightarrow S) \qquad (4)$$

$$I_{RULE}(R_X \text{ U } R_Y \rightarrow S)=\text{Contribution}(R_X \text{ U } R_Y \rightarrow S)/\text{LIFT}(R_Y \rightarrow S) \qquad (5)$$

Where, I_{RULE} indicates whether the contribution of R_X (or R_Y) to the occurrence of *S* increases, while considering R_Y (or R_X) as a precondition to the rule. To get more information, we also generate pair patterns [5], cluster pattern, incremental pair-pattern and incremental cluster-patterns, and their respective contribution and interestingness matrices. In case of pair pattern, two atomic rules are taken to form a pair-pattern if and only if the two atomic rules have at least one common data-item in their antecedent parts and after removing those common data-item/data-items from the atomic rules the antecedent parts of none of the atomic rules should be null. In case of incremental pair-pattern, we actually remove the common data-item/data-items from the pair-patterns and consider the common data-items as a pre-condition. In case of cluster pattern formation, we try to include as much as rules in a single cluster on the basis of the common data-item/data-items in their antecedent parts and also take care of the fact that after removing the common data-item/data-items from the antecedent parts of the respective rules, the antecedent part of the rules should not be empty or null. In case of incremental cluster-patterns, we actually remove the common data-

item/data-items from the respective rules in a cluster and consider the common data-item/data-items as a precondition for that particular cluster of rules.

Multi-method Mining Approach: In this approach, first the association rules by using FP-tree algorithm[2] and [15] are generated and then make the Probabilistic Frequent Item-set [13], to be used by those association rules during training phase and then classify the testing data-set by Probabilistic Frequent Item-set during testing phase. In this approach, first the association rules by using FP-tree algorithm[3] and [6] are generated and then make the Probabilistic Frequent Item-set [12], to be used by those association rules during training phase and then classify the testing data-set by Probabilistic Frequent Item-set during testing phase. With this approach it will extract the threshold-based PFIs from large databases. Probability model is used to obtain the probability mass function, and also support probability mass function is calculated which will generate the results based on probability approximated values.

IV. RESULTS

A. Data set

We have taken two bunches of data- set, one from yahoo finance and another from UCI-Machine Learning Repository, Census Income data set which is also named as "Adult" for our project work. This data-set have 14 attributes, some of them are discrete attributes, while others are continuous attributes. The dataset for Yahoo Finance consist of the various number (Bunch) of Companies.

B. Result set

The default value of minimum probability is set to 0.1 and minimum support value is set to 0.5. We first present the results for mining the data sets using FP tree and then applying it along with PFI testing method. Then the attribute analysis graph is shown. In Fig. 2 and 3 the time measurement is taken in milliseconds.

Fig. 2 and 3 shows that using single method FP Tree is less efficient in generating actionable patterns compare to multi-method approach such as PFI with FP Tree. Fig. 3 shows the impact of probabilistic measure towards generating frequent item-set which also considers the boundary values which has likeliness towards the actionable patterns.

Fig. 4 describes the result for uncertainty in attributes data. We have the result of combined mining with analysis of the attributes when 1, 2 or 3 attributes are in common in both data sets respectively.

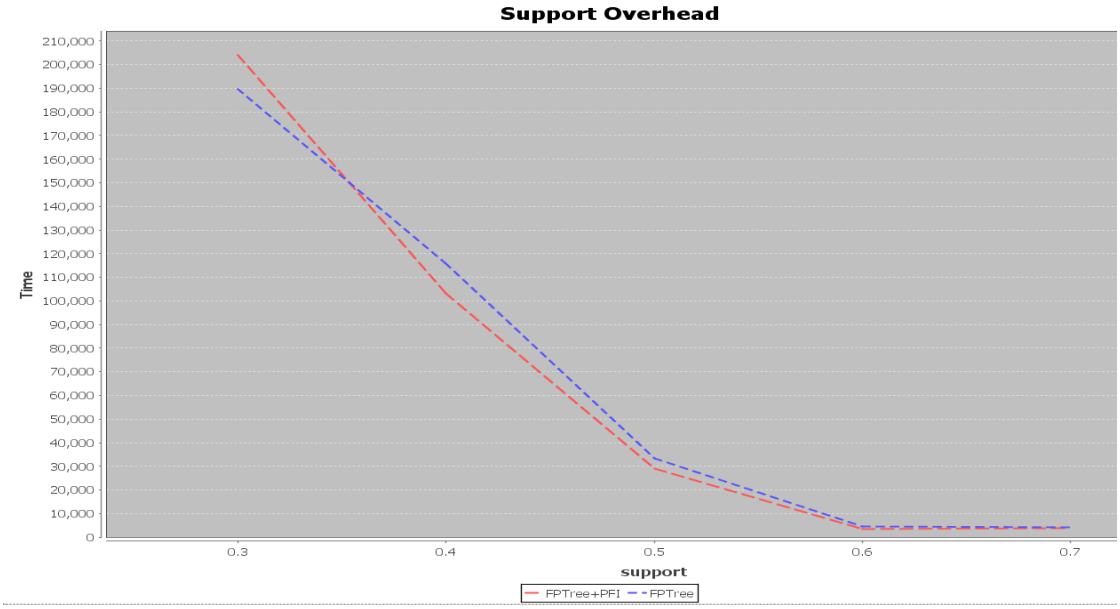

Fig. 2 Support Vs. Execution Time for Generating Actionable Patterns using Adult and Bank Marketing data sets

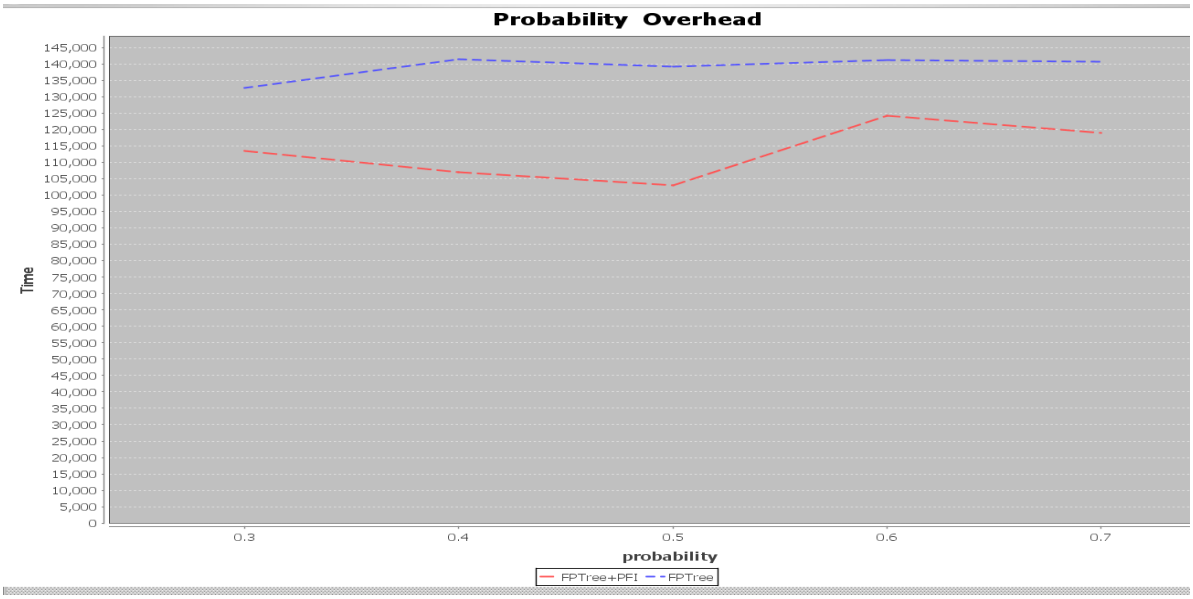

Fig. 3 Probability Vs. Execution Time for Generating Actionable Patterns using Adult and Bank Marketing data sets

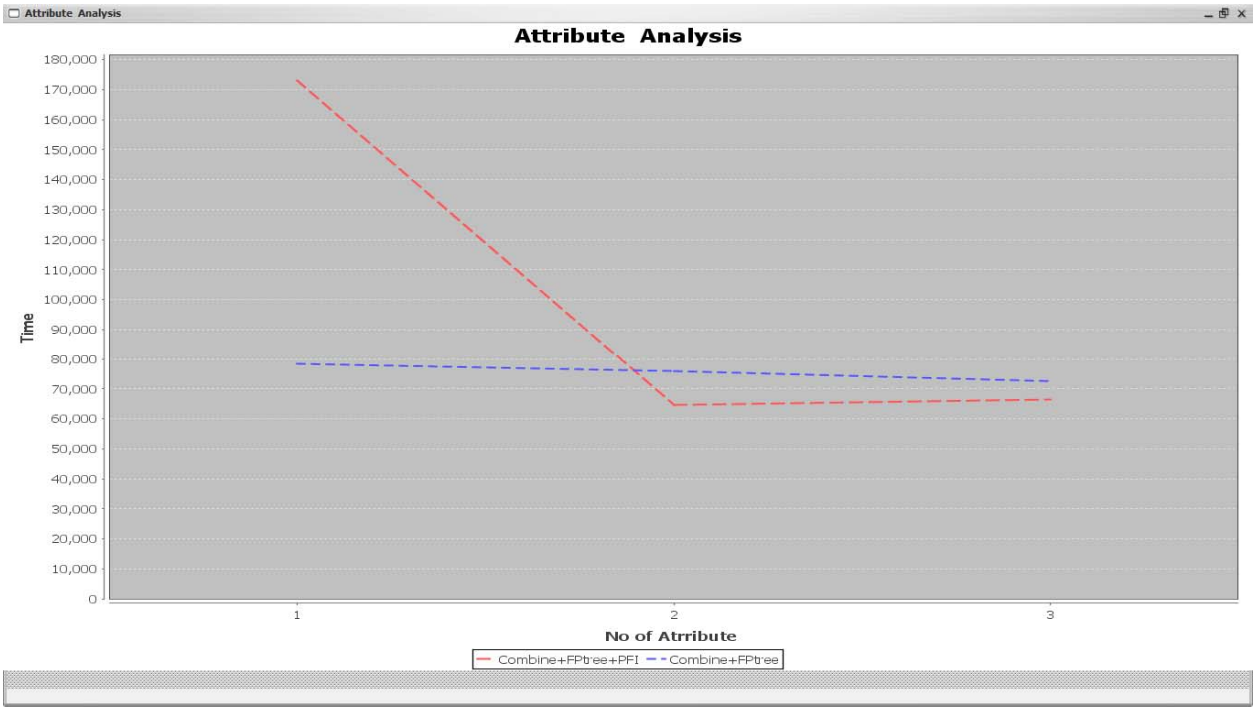

Fig. 4 Combined Mining Attribute Analysis

V. CONCLUSION

In this paper we propose an approach to extract threshold PFIs from large and heterogeneous data sources. The support have impact on run-time classification percentage by Probabilistic Frequent Item set (PFI), while probability of not having effect of confidence neither on run-time nor on classification percentage. The combined patterns identified here are more informative, actionable and impact-oriented as compared to any single patterns identified by traditional methods like FP-growth, etc. We can have such frameworks, which are flexible and customizable for handling a large amount of complex data, for which data sampling and table joining may not be acceptable. Further developments are possible to handle large and multiple sources of data available in industry projects for government, stock market, insurance, banking, etc. with effective paradigms and interestingness measures.

ACKNOWLEDGEMENT

Author takes this opportunity to thank, guide and HOD, Dr. A. B. Bagwan, for his continuous encouragement and valuable inputs and Dr. P. K. Deshmukh for his impeccable suggestions. Authors also thank, Dr. D. S. Bormane, Principal, RSCOE and all teaching, nonteaching staff members for their support.

References

1. L. Cao, H. Zang, Y. Zhao, Dan Luo, C. Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data" IEEE Transactions on Systems, Man, And CyberneticsPart B Cybernetics. Vol. 41, NO. 3, Year 2012.
2. Han and Kamber (2006), "Data Mining Concepts And Techniques", 2nd ed., United State of America.
3. 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.
4. Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang, "Mining for combined association rules on multiple datasets," in Proc. DDDM, 2007, pp.18–23.
5. 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.
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, pp. 1299–1312, Sep. 2010.
7. H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Combined association rule mining," in Proc. PAKDD, 2008, pp. 1069–1074.
8. M. Plasse, N. Niang, G. Saporta, A. Villeminot, and L. Leblond, "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set," Comput. Statist. Data Anal., vol. 52, no. 1, pp. 596-613, Sep. 2007.
9. X. Yin, J. Han, J. Yang, and P. S. Yu, "Efficient classification across multiple database relations: A CrossMine approach," IEEE Trans. Knowl. Data Eng., vol. 18, no. 6, pp.770–783, Jun. 2006.
10. 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.
11. B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in Proc. 4th Int. Conf. Knowl. Discov. Data Mining (KDD), 1998, pp. 80–86.
12. Xenofontas Dimitropoulos, Paul Hurley, Andreas Kind "Probabilistic Lossy Counting: An efficient algorithm for finding heavy hitters"
13. Liang Wang, David Wai-Lok Cheung, Reynold Cheng, Member, IEEE, Sau Dan Lee, and Xuan S. Yang, "Efficient Mining of Frequent Item Sets on Large Uncertain Databases" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 12, December 2012
14. Longbing Cao "Combined Mining: Analyzing Object and Pattern Relations for Discovering Actionable Complex Patterns"
15. Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao, " Mining Frequent Patterns Without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, 8, 53–87, 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.
16. B. Liu, W. Hsu and Y. Ma (1999), "Pruning and summarizing the discovered associations," in Proc. KDD, pp. 125–134.
17. Kargupta , B. Park, D. Hershbereger, E. Johnson and H. (1999), 'Collective data mining: A new perspective toward distributed data mining'. Accepted in the Advances in Distributed Data Mining, Eds: Hillol Kargupta and Philip Chan, AAAI/MIT Press (1999).
18. Jianyong Wang and George Karypis(2005), "HARMONY: Efficiently Mining the Best Rules for classification" SIAM International Conference on Data Mining, pp. 205--216, 2005.
19. A. Archana and T. Mathavi Parvathi "Knowledge Extraction By Combined Mining" International Conference on Computer Science and Engineering (CSE) 7th April 2013, Bangalore, ISBN: 978-93-82208-83-9.

AUTHOR(S) PROFILE

Ms. Seema A. Shabadi, B. E. (Comp. Sci. and Engg.)-2003, Aurangabad University, M.E.(Comp. Engg.) Pursuing. She is currently working as a lecturer in Comp. Engg. Dept. of RSCOE, Tathawade, Pune.

Dr. A. B. Bagwan, (M'92), B.Tech.(Electrical Engg.)-1973, IIT Kanpur, M.Tech.(Comp. Sci.)-1977, IIT Madras, Ph.D.(Comp. Sci. and Engg.)-2012 Bundelkhand University. He is currently working as a Head of Department in RSCOE, Tathwade, Pune. Dr. A. B. Bagwan is a member of ISTE since 1992.

