Knowledge Extraction Using Combined Mining Approach With Parallel Processing

Sharayu Pawar
ME Student, Dept. of Computer Engg.
MET's BKC Institute of Engineering
University of Pune, India.

Madhuri Bhalekar
Assistant professor, Dept. of Computer Engineering, MET's BKC IOE,
University of Pune, India.

Dr. M. U. Kharat
Professor, Department of Computer Engineering, MET's BKC IOE,
University of Pune, India.

Abstract**—** Data Mining is rapidly developing field with algorithms and methodologies still being invented. Thus there is still a gap between available methodologies of Data Mining and tools for data mining to extract domain specific knowledge. The traditional data mining methods can be used on single source of data, and they do not suit well to large and complex data. Also these traditional methods can perform mining on homogeneous kind of data only. For data environment, where Large and complex data from multiple sources and even data containing multiple heterogeneous features is involved, combined mining approach is essential. This approach is can extraction actionable knowledge from large data that can be useful for making business decision. Outcome of the combined mining approach is represented as Combined Patterns, which shows overall analysis of data from different perspectives.

Keywords- *Data Mining, Heterogeneous Data, Complex Data, Combined Mining, Knowledge Discovery from data, actionable knowledge discovery, multi-source combine mining*

____*****____

I. Introduction

Data mining is the process of extracting knowledge from large amount of data that is hidden and previously unknown. It is the set of rules and methods used to find new hidden patterns in data. The data mining can give answers to many business problems that can help in taking major decisions such as

- What marketing strategy to employ?
- Which customers to focus for specific product?
- How budgeting should be done to departments?
- Deciding likelihood that a certain customer will default or pay back a schedule.
- Give appropriate medical diagnosis for a particular patient.

These types of queries can be handled easily if the data mining is done to extract the useful information hidden in the data. [7] But, in case of mining complex types of data certain challenges are faced by traditional data mining methods. Same is the case with handling of extremely large datasets, generated as a result of WWW, Organizations, Governments, Institutions and many more. But the data is growing at a faster rate than the techniques to mine that data.

Following are few limitations of the traditional data mining methods:

a) Joining multiple relational tables, as traditional data mining methods work on single source only.
b) Mining data containing heterogeneous features is not possible by a single traditional data mining method.
c) Since generated outcome can be huge for a sufficiently large dataset, Post analysis on mining is required.
d) Since traditional data mining methods generate only technical interesting knowledge, extracting the actionable knowledge may require Involvement of multiple mining methods.

Combined mining is a solution that comprise of traditional data mining methods, so as to perform "Combined Mining" on multiple sources of data containing multiple features. Large and complex data is efficiently handled by the Combine Mining approach. This approach delivers combined patterns which are more domain driven, business oriented and which provide in depth analysis of data. These patterns are more informative, comprehensive and are actionable, useful in decision making process. The combined mining approach is generalized concept and an extension to traditional mining methods.

II. Previous Work

Backward references for Combined Mining approach can be given as efforts to overcome the limitations of traditional data mining approach while mining large and complex data. Many attempts were made to make traditional data mining process more domain driven and all pervasive. These approaches include

a) Combined Association Rule Mining
b) Post Analysis Based Approach
c) Uniform Interestingness based approach
d) Direct Mining Advancements

Authors H. Zhang et al. proposed the concept of Combined Association Rule Mining in [4].The process of finding combined association rule is a two- fold. At first, rules with frequent itemset are found, and then combined association rules are extracted from them using interestingness measure named conditional lift.

In work proposed by authors in [2] PA-AKD Framework is given which is based on the concept of post analysis of generated rules, the term is Post Analysis Based Actionable

Knowledge Discovery. This approach is also two step approach, namely rule extraction and result refinement. In [3], Liu et al. proposed a pruning and summarization rules for post analysis of data. According to proposed work, rules are classified as Direction Setting and non-Direction setting rules. Chi-square ($\chi 2$) test is used as the pruning criterion.

In [2] authors H. Zhang Et.al, proposed UI-AKD framework, which is a Uniform Interestingness based approach. In this approach, Uniform Interestingness metrics is used, that can be used to extract patterns with both technical and business specification

III. PROPOSED APPROACH

A. *Problem Definition*

Many organizations today store their data at a number of locations, for safety and reliability purpose. Mining such data is not possible by traditional data mining methods.

Even if mining is performed, the knowledge generated would be of technical interestingness (TI) only. Such kind of extracted knowledge is of no use to business, because frequently occurring patterns might be common sense, and are not expected to be calculated mathematically.

For taking serious business related decisions, and for making a proper marketing strategy, knowledge with a high level of business interestingness (BI) is required. Such knowledge can be extracted from large and multiple datasets either by joining the tables, or by applying multiple methods. Both these approaches are difficult or even impossible to implement in some cases.

So for this purpose a flexible and generalized approach of mining is needed to be developed which can perform data mining in such environments. The approach need to address following issues in traditional data mining methods-

- Mining data from multiple sources: for covering many aspects of the business
- Mining data containing multiple features: for reflecting characteristics of business
- Mining data using multiple mining methodologies: for reflecting in depth nature of data and also provides advantages of various methods used [1].

All these limitations can be overcome by making use of CM approach, which mines data using existing method, then combining the outcomes and then applying multiple interestingness measures to those combined patterns, resulting in generation of highly actionable, business oriented knowledge.

Combined Mining thus is a new approach for performing data mining on the heterogeneous data where following aspects are considered: Mining data from multiple sources, Mining data containing multiple features, Mining data using multiple mining methodologies. Also metrics are of multiple interestingness are applied to the generated patterns so as to verify the significance of generated patterns, in terms of business interestingness (BI).

B. *Basic Process of Combined Mining*

The main function of Combined Mining approach is to generate business oriented knowledge from complex data, containing multiple features.

Thus, a basic process for Combined Mining can be given as: Patterns $P_{n,m,l}$ are the generated combined patterns by using data mining method R_l performed on features F_k from a data set D_k with the use of interestingness $I_{m,l}$ [1]

$$P_{n,m,l} : R_l (F_k) \rightarrow I_{m,l}$$

$$\text{where } n = 1, ..N;\ m = 1, ..M;\ l = 1, ..L.$$

Here, Data set D_k Represents the data sources or partitioned subsets. Feature set F_k refer the features used for performing mining on D_k. Method set R_l, is a data mining method set to be performed on Dk. Interestingness set $I_{m,l}$ is set of interestingness metrics. Lastly, Pattern set $P_{n,m,l}$ is the pattern set (atomic pattern set) generated as a result of data mining method R_l using interestingness $I_{m,l}$ [1].

The input to the Combined Mining system can be a single large dataset or even multiple datasets. In case of single large dataset, on which mining can be difficult to implement, partitioning can be performed by user, as per business requirements. Also, the choice of which features to be included in each partition is left to user [8].

For each partition, atomic patterns are generated and then merging is done to generate combined patterns. The outcome of the system is delivered as Combined Patterns, generated both sequentially and parallelly. This approach utilizes the existing methods of data mining and combines them so as to adapt to before mentioned needs [8]. The architecture of combined mining is given in figure 1.

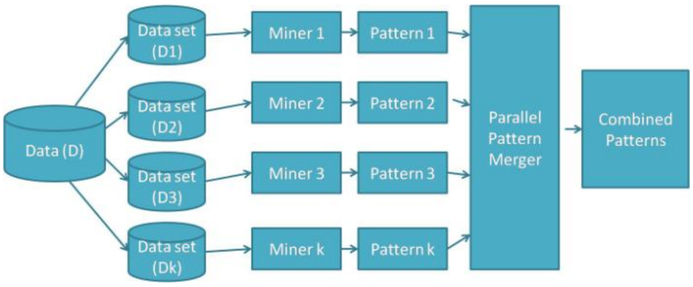


Figure 1. System Architecture for Combined Mining Approach

The Combined Mining Process as proposed by us is multistep one, as compared to traditional data mining methods, which are single step. The steps for combined mining approach are given below [8]

1) Load Database: In the first phase of Combined Mining, the data is loaded from one or more sources..

If the data is too large to handle, then Partitioning can also be performed in this phase.

2) Feature Selection: Here Selection of Business Related Features is done, Based on domain knowledge, Goal definition etc. for the first dataset. For Next Dataset, same features can be kept or other required features can be added, as per findings from previous dataset.

3) Pattern Generation: This step corresponds to mining atomic patterns $P_{n,m,l}$ from individual dataset Dk. Steps 2 and 3 are repeated for k times, to generate k atomic pattern sets.

4) Pattern Merging: Here Merging of atomic pattern sets into combined pattern set P = G(P_k) is done. Using pattern merging method Gk, suitable for a particular business problem.

5) Patterns Visualization: In this step Combined Patterns are presented as deliverables.

C. *Parallel Algorithm for Combined Mining:*

Algorithm for the process of Combined Mining as supported by our parallel merger architecture is as given below. First phase of algorithm can be implemented on a common CPU Architecture, whereas second phase of algorithm will be implemented on a GPU [8].

```
Algorithm: Combined Mining
Input: Multiple Data sets with partitioning.
Output: a complete set of combined patterns.

Phase I: CPU Processing
    1.  Load Data set D_i from disk.
            a.  Select attributes and mining algorithm set by
                user to perform mining on data set D_i.
            b.  Perform mining on given data set Di and get
                atomic patterns.
    2.  Transfer the atomic patterns from all the Data sets to
            GPU memory.

Phase II:  GPU Processing
Perform Parallel merging of all the atomic patterns.
On each kernel/work-item/thread (i) do,
    1.  Scan i^th atomic pattern from shared memory
    2.  for each item in  i^th atomic pattern do,
            Scan the memory for merging given atomic
            pattern to get one of the following.
             a. Pair pattern.
             b. Cluster pattern.
             c. Incremental Pair Pattern.
             d. Incremental Cluster pattern.
        Endfor
```

D. *Combined Patterns Representation*

The outcome of Combined Mining is represented in the form of combined patterns, covering all aspects of business. These combined patterns can be one of the following forms: [1]

1) Pair Pattern:
For combined mining, a pair pattern is in the form of

$$P = \left\{ \begin{array}{l} X_1 \rightarrow T_1 \\ X_2 \rightarrow T_2 \end{array} \right.$$

Interestingness of a combined association rule pair is calculated using

$$\begin{array}{lll} \text{Ipair}() = & |\text{Conf}(P_1) - \text{Conf}(P_2)| & \text{if } T_1 = T_2 \\ & \surd\text{Conf}(P_1)\text{Conf}(P_2) & \text{if } T_1; T_2 \text{ are contrary} \\ & 0 & \text{otherwise} \end{array}$$

2) Cluster Pattern:
If there are k atomic patterns $X_i \rightarrow T_i$, (i = 1, . . . , k), k ≥ 3, and $X_1 \cap X_2 \cap \ldots \cap X_k = X_p$, a cluster pattern (P) is in the form of

$$P = \left\{ \begin{array}{l} X_1 \rightarrow T_1 \\ \cdot\ \cdot\ \cdot \\ X_k \rightarrow T_k \end{array} \right.$$

Interestingness of a combined association rule cluster is calculated as $I_{cluster}(P)$ = maxPi;P j∈C;i≠j Ipair(P_i, P_j)

3) Incremental Pair Pattern:
An incremental pair pattern is a special pair of combined patterns as follows:

$$P = \left\{ \begin{array}{l} X_p \rightarrow T_1 \\ X_p \wedge X_e \rightarrow T_2 \end{array} \right.$$

where $X_p \neq \Phi$, $X_e \neq \Phi$, and $X_p \cap X_e \neq \Phi$.

The interestingness of Incremental Pair Pattern is measured using another interestingness measure called Cps() as proposed by [1]

4) Incremental Cluster Pattern:
An incremental cluster sequence is a special cluster of combined patterns with additional items appending to every previously adjacent constituent patterns. An example is

$$\mathcal{P}: \begin{cases} X_{\mathrm{p}} \rightarrow T_1 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \rightarrow T_2 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \rightarrow T_3 \\ \cdots \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \wedge \cdots \wedge X_{\mathrm{e},k-1} \rightarrow T_k \end{cases}$$

The interestingness of Incremental Cluster Pattern is measured using another interestingness measure called as Impact()

IV. Implementation Strategy

A. Implementation Details

The Combined mining system works by taking as input one or more datasets. For a single large dataset, an option for partitioning is provided, which partitions data into a number of subdatasets. After that, atomic patterns are generated on each dataset. For generating these patterns, user is given the choice to select desired features, and also to select minimum threshold value. He can select those features only, which may help him in decision making. Now, after the atomic pattern generation, the patterns are merged both sequentially and parallely. The output of system is delivered in the form of combined patterns. Speedup achieved by parallel merging approach, as compared to sequential is shown by means of graph. For parallel processing of the algorithm, coding is done according to GPU hardware architecture. OpenCL programming language is used for the coding purpose [8].

B. Parallel Pattern Merger

As proposed by [1], the combined patterns are termed as Pair Pattern, Cluster Pattern, Incremental Pair Pattern, and Incremental Cluster Pattern. These patterns are delivered as result of combined mining, generated from large number of atomic patterns by taking into consideration both technical interestingness and business interestingness. But while merging the patterns to generate combined patterns, a penalty in performance is faced by system, if the processing is done sequentially. To overcome this limitation incurred by existing implementations of combined mining approach, we have proposed a new parallel merger for the purpose. [8]

This parallel pattern merger can work on parallel processing CPU architecture such as SIMD and also on GPU, making use of GPGPU concept [8].

V. Results Generated

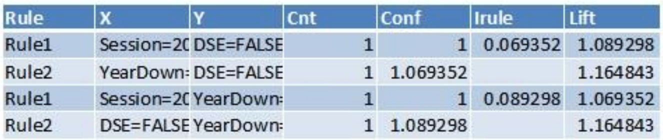
A. Dataset Used:

The dataset on which our system is tested is a dummy dataset from ERP system of an educational institute, named MET’s Bhujbal Knowledge City, Nasik. The data consist of multiple features such as Session, Institute, DSE, year, Branch, Provisional, YearDown, Scholarship, Sco Form Received, Hostel. Etc.

Along with this dataset, we have checked the system on few more datasets, which include German Credit Data, Land Registry data, and Census Income Data.

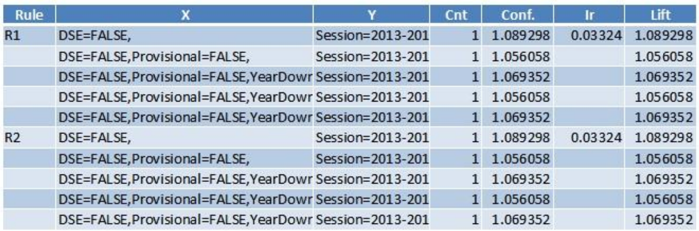
B. Result Generated on MET Data:

When testing is done on MET Data, by giving 60% support and selecting 8 features in two datasets, the pair patterns and cluster patterns generated are as shown in table. All the values of Interestingness measures are also reflected in Table, which include Lift, Confidence, and I_{rule} .



Rule	X	Y	Cnt	Conf	Irule	Lift
Rule1	Session=20	DSE=FALSE	1	1	0.069352	1.089298
Rule2	YearDown=	DSE=FALSE	1	1.069352		1.164843
Rule1	Session=20	YearDown=	1	1	0.089298	1.069352
Rule2	DSE=FALSE	YearDown=	1	1.089298		1.164843

Table 1. A snapshot of Pair Pattern



Rule	X	Y	Cnt	Conf.	Ir	Lift
R1	DSE=FALSE,	Session=2013-201	1	1.089298	0.03324	1.089298
	DSE=FALSE,Provisional=FALSE,	Session=2013-201	1	1.056058		1.056058
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.069352		1.069352
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.056058		1.056058
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.069352		1.069352
R2	DSE=FALSE,	Session=2013-201	1	1.089298	0.03324	1.089298
	DSE=FALSE,Provisional=FALSE,	Session=2013-201	1	1.056058		1.056058
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.069352		1.069352
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.056058		1.056058
	DSE=FALSE,Provisional=FALSE,YearDowr	Session=2013-201	1	1.069352		1.069352

Table 2. A snapshot of Cluster Pattern

C. Result Analysis

The comparison for number of delivered patterns by traditional data mining system and Combined Mining System is given in figure2. The analysis of both approaches shows significant reduction in number of patterns generated by our Parallel processing Combined Mining approach, as compared to traditional data mining approaches. This makes it easier for user to understand the mined knowledge, as less number of patterns are needed to be analyzed.

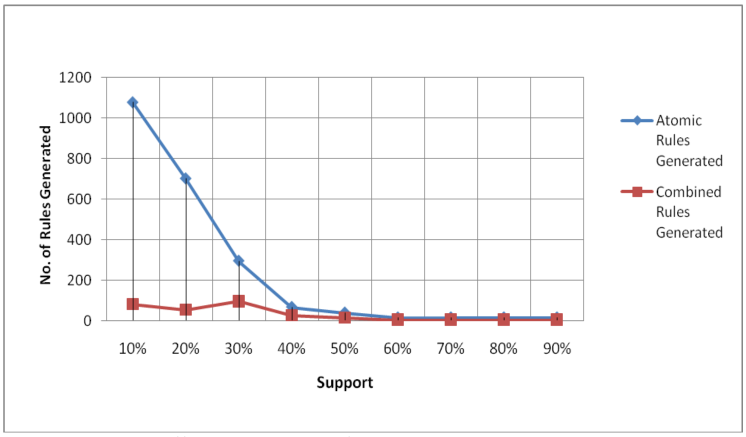


Figure 2. Efficiency Analysis of Combined Mining Approach

Speedup achieved by our parallel approach for Combined Mining is shown by table 3. It is found to outperform the existing system for Combined Mining as given in [1], which is based on sequential approach.

Min_Support	Sequential Time	Parallel Time
10%	65	6
20%	37	6
30%	18	6

40%	12	6
50%	16	6
60%	12	7
70%	6	6
80%	5	6
90%	5	6

Table 3. Speedup Achieved by our Parallel Merging Approach as compared to sequential Approach

The analysis of table shows that when lare number of patterns are generated, as a result of very small support threshold, there is significant reduction in time required by parallel approach as compared to sequential one. At a certain level, both approaches give same complexity, which is 70% support in this case. Also at a level where only a small number of patterns are generated i.e. when highest support threshold is provided, the time rquired by parallel approach may be greater due to large amount of time consumed in only memory transfer from CPU to GPU. But still the overall time coplexity of Parallel approach remains in same range which is sufficiently low, and which is not the case with sequential approach. The same analysis is shown in graph 3.

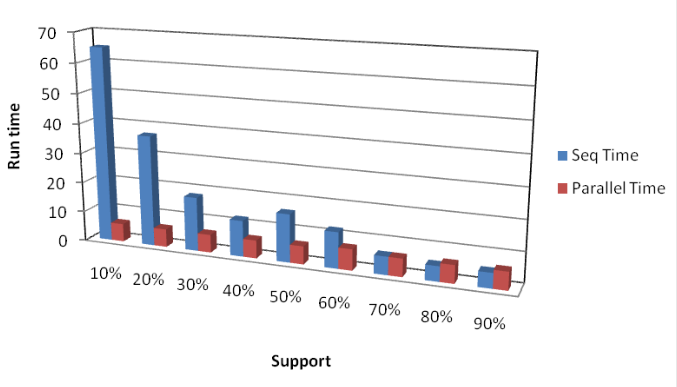


Figure 3. Performance Analysis of Combined Mining Approach in terms of speedup achieved.

VI. Result Analysis

Combined Mining Technique delivers combined patterns which are more domain driven, business oriented and which provide in depth analysis of data. These patterns are more informative, comprehensive and are actionable, i.e. they are useful in decision making process of critical business related decisions [8]. This makes combined mining approach more advantageous than traditional data mining methods, which work on single source of data, and mostly homogeneous data. The combined mining approach is flexible and can be applied as per business requirements. This approach is implemented by applying traditional data mining methods to multiple sources, and by mining multi-feature data and even by using multiple methods for mining data [8].

The system proposed by us, involve generation of atomic patterns by using multisource, multifeature and multimethod approach, followed by the process of pattern merging, which we have done by parallel processing methodology [8].

This system, as compared to existing systems of combined mining, performs better in terms of time complexity, since the existing systems follows sequential approach, as compared to our parallel one.

VII. Conclusions

The combined mining approach is invented to overcome the limitations of traditional data mining algorithms, while handling large and complex data. The combined mining system as developed by us is capable of generating knowledge, which is business problem oriented and actionable, by mining data that can be heterogeneous and distributed in nature. Our system uses a parallel approach for performing combined mining and thus outperforms existing systems in terms of time complexity.

References

[1] 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:, VOL. 41, NO. 3, JUNE 2011

[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] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," in Proc. KDD, 1999, pp. 125–134.

[4] H. Zhang, Y. Zhao, L. Cao, and C. Zhang, "Combined association rule mining," in Proc. PAKDD, 2008, pp. 1069–1074

[5] Y. Zhao, H. Zhang, L. Cao, C. Zhang, and H. Bohlscheid, "Combined pattern mining: From learned rules to actionable knowledge", in Proc. AI,200q8, pp. 393–403.

[6] Jennifer Widomy Brian Lenty, Arun Swamix. "Clustering association rules". *1997 IEEE*, pages 220–231.

[7] Tamanna Sehgal Gaurav Gupta, Geetika Hans. "Applications trends in data mining."

[8] Sharayu Pawar, M. A. Bhalekar, and M. U. Kharat, "Knowledge Discovery using Combined Mining Approach", in Procc. Of The Third Post graduate Symposium of Computer Engineering, cPGCON-2014.
.