Combined Mining And Actionable Pattern Discovery Using DDID-PD Framework: A Review

Mrs. Suvarna R. Bhagwat
JSCOE, Pune, Maharashtra.

Abstract

In combined mining, the word "combined" principally refers to either one or more of the following aspects on demand, 1) Combination of multiple data sources, 2) Combination of multiple features, & 3) Combination of multiple methods.

The outcomes of combined mining are combined patterns, which are actually the patterns evaluated from heterogeneous sources. Such patterns reflect characteristics of the every source from which they are extracted as they contain features from various sources. One can also use more than one method such as association rule mining, clustering, classification, prediction etc. for mining same data. For example generated frequent patterns can be classified in order to derive more informative knowledge. Resultant combined patterns surely have a complete essence of data by taking advantage of different methods. But sometimes such patterns mined from heterogeneous sources, by multiple methods may be only of technical interest i.e. useful to the miner and not to user.

In order to satisfy the need of particular business application these patterns need to be treated by various interestingness metrics so that they reveal importance as well as concerns to the required perspective. Such patterns are known as actionable patterns. Interestingness matrices need to be developed by taking into account many aspects such technical performance, business performance, domain knowledge, end user experience as well as organizational & social factors.

1. Introduction

Data mining has already been widely used in many areas such as public services, telecom, share market, shopping malls, health care and many more. In new era, what developers are concerned about are the various types of data sources involved in applications. Now days the data sources involve heterogeneous data for example Transactional data, XML documents, Text Files etc. Also the tractional data soursce may involve multiple features. To handle such multifeatured, heterogeneous data sources the process of mining needed to be generalized. The combined mining process does the same. Following are some scenarios which elaborate different approaches of combined mining.

1. The transactional data sources involved in data mining applications may have multiple features. Combined mining selects the features from all sources which has more importance, and incorporate them into resultant patterns. Such patterns are known as combined patterns (as they contain multiple features from heterogeneous data sources.)
2. Sometimes the data to be mined can be distributed or volume of data can be so large that it is impossible to scan the whole data. Combined mining scans each data source separately and then combine the generated patterns.
3. As known, there are many methods of data mining for example association rule mining, Classification, Clustering, Summarization, Prediction, etc. But many times outcome of a single method may not be useful in required perspective. Combined mining make use of multiple methods to generate patterns which reveal real meaning of data by taking up the advantages of multiple method.
4. The patterns generated by using multiple features from heterogeneous sources as well as by multiple method of mining are clearly a treat to miner. But end user may not be at all interested in this knowledge as it may not serve to its perception. In order to make discovered patterns ready to use in process of decision making they need to be treated by various interestingness matrices. The process of making the discovered patterns actionable should be interactive, iterative and should indulge in-depth knowledge of domain.

The remaining part of this paper is organized as follows. In section II brief idea of multisource mining has been reviewed. Section III and IV helps to elaborate multifeature and multimethod combined mining approaches. In section V the DDID-PD Framework for generation of actionable patterns has been reviewed. The paper is concluded in section VI.

2. Multisource Combined Mining

As discussed in introduction, in this era input of data mining application comes from various data

sorces. These data sources can be heterogeneous, complex as well as huge in size. Rather than combining this data into a single homogeneous form (which is generaly done by table-joining operations) and then operated upon by mining algorithms, it is preferable to combine discovered patterns from various heterogeneous data sources[1].

Figure 1 shows the steps to be taken to implement multisource combined mining appraoch.

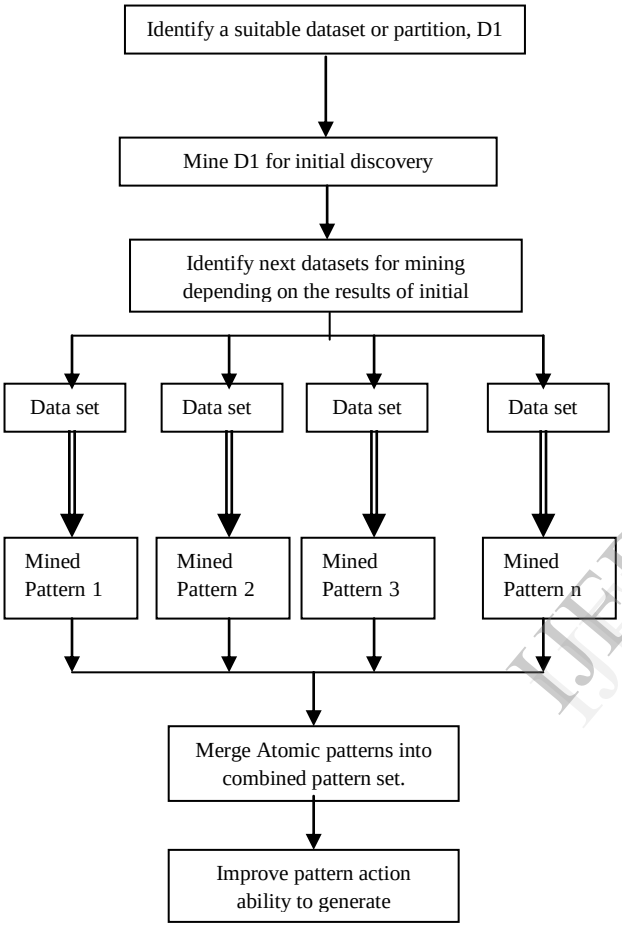

Figure 1. Multisource Combined Mining

3. Multifeature Combined Mining

Multiple data sources with homogeneous patterns are easy to handle, but data sources with multiple fetaure-set needs to be handled with special management. In multifeature combined mining, atomic patterns (generated from a single source) are merged together to form combined patterns which are often more informative. Following are some types of combined patterns[1].

- Pair patterns can be formed by combining two atomic patterns. They are of form, $\{A_1 \rightarrow B_1, A_2 \rightarrow B_2\}$ where A_1 and A_2 are same but B_1 and B_2 are different or vise versa. New measure $I_{pair,}$ is used to measure the interestingness of pair pattern.
- Cluster patterns, are formed by organizing many similar or related atomic or pair patterns together. They take the form, $\{A_1 \rightarrow B_1, A_2 \rightarrow B_{2,} A_N \rightarrow B_N\}$. Measure $I_{cluster}$ defines how interetsed is the cluster of patterns.
- Some patterns can take form of extension of other patterns. For example, $\{A_1 \cup B_1 \rightarrow C_1\}$ can be thought of an extension of $\{A_1 \rightarrow C_1\}$. Such patterns are combined to form incremental pair patterns.
- Many patterns which are extension of one another can be grouped together to form Incremental cluster patterns, as $\{A_1 \rightarrow Z_{1,} A_1 \cup B_1 \rightarrow Z_{1,} A_1 \cup B_1 \cup C_1 \rightarrow Z_{1...}\}$.

In Figure 2, the procedure to discover multifreature combined patterns is given in simple steps.

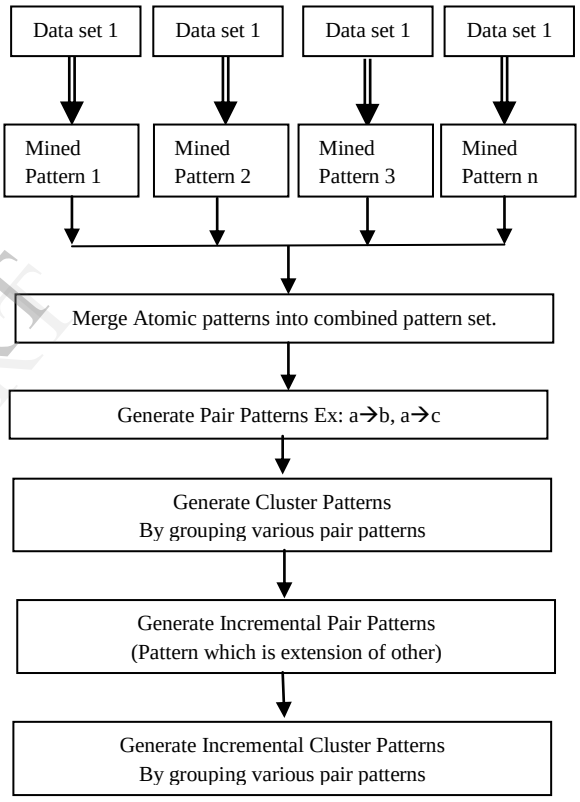

Figure 2. Multifeature Combined Mining

4. Multimethod Combined Mining

In many situations the patterns discovered by a paricular method do not serve to user's perpsective. Here one finds need of using more than one mining methods in order to discover more informative patterns. Multiple methods can be used parallely, serialy or in closed loop fashion[1].

4.1 Parallel Multimethod Mining

In this approach various mining techniques are used parallely on different or same data sets.

1. Initially different, independent data sources are mined using different data mining techniques, to find out respective atomic pattern sets.
2. Then these patterns are merged together by merging method.

4.2 Serial Multimethod Mining

Here various data mining techniques are used one by one. Outcome of one technique is treated by another technique to discover in depth knowledge. This method works as follows.

1. Initialy the datasource is mined using a suitable method to obtain pattern set, say P_1.
2. After studing the initial pattern set, next suitable method is selected and P_1 is again mined using it to discover next level pattern P_2.
3. Various techniques are applied according to domain knowlwdge and output requirement.

4.3 Closed loop Multimethod Mining

This appraoch takes into account the impact of one technique on other. Practicaly feedback from next method can be used to refine the results of previous method. So one can easily observe that in closed loop approach, the decision (whether the pattern is interested or not) does not depent upon a particular method but also on other methods used in the process. Closed loop mining is carried out as follows.

1. Initial pattern set P_1 is formed by following the process of serial multimethod mining.At the end of this step some samples may not be identified. This is due to the limitations and conditions applied on various mining techniques.
2. The patterns in P1 are checked for their validity. Some samples may not be valid to patterns. A separate data set is formed using such exceptional samples, say D_x. This dataset is again mined by mutliple method to discover new pattern P_2.
3. Process of step 2 repeted as many times required by miner to discover patterns $P_1,P_2,\ldots\ldots P_k$.
4. All patterns are then merged to form combined pattern.

5. DDID-PD Framework

Many times it is observed that the patterns which are helpful in real life problem solving are hidden among large quantity of complex data discovered during mining. Or sometimes the discovered patterns happen to be useful to technical individual than to its real user. Morever many discovered interesting patterns are often can not be applied directly in real life but they should be integrated with business rules or social constraints.

DDID-PD framework is domain driven, in-depth pattern discovery process which allows interactive, iterative actionable pattern discovery in domain specific perspective.

5.1 Actionability of Patterns

In DDID-PD framework actionability of patterns is measured in subjective as well as objective perspectives. Also the discovered patterns should be both of technical & business interest [3].

5.1.1. Technically interetsed pattern Such pattern is said to be dependent on certain technical measure defined for a particular mining technique. Technical interestingness is measured in terms of technical_objective and technical_subjective measures.

1. Technical_objective measures are set of criteria which decide whether given pattern is interested or not. For example in case of association rule mining a pattern is said to be interested if it satisfies minimum *support*, and minimum *confidence* measures.
2. Technical_subjective measures help to rank up to what extent the discovered pattern is useful to a specific user's need.

5.1.2 Business interested pattern Whether a pattern is interested in business or not, is need to be decided using various aspects viz. economical, social, analytical and personal perspectives. Just like Technical interestingness, Business Interestingness is also measured in terms of business_objective and business_subjective measures.

1. Business_objective measures are criterias depending upon economical, social, or a particular business person's perspective. For example in market basket analysis, *profit* can be used to determine business potantial of a mined pattern. If a predecided profit is achieved using minned pattern then that pattern is said to be interested in terms of business_objective measure.
2. Business_subjective measures are said tobe psychoanalytical measures. For example, in case of stock trading analysis experienced brokers' thinking can be used as business_subjective measure.

A pattern is said to be *actionable* if it satisfies both technical as well as business interestingness measures. To discover such pattern, miner and user must work collaboratively.

5.2 DDID-PD Process Model

According to this framework the process of data mining has following steps. The sequence of these steps can be flexibly changed. Some steps can be moved back or forth, some steps can be repeated, or some can be dropped according to need in real life application. The outline of data mining process is shown in figure 3. The blocks highlighted using double bordered boxes are particularly from DDID-PD framework.

To elaborate more about DDID-PD process, following are reviewed some features. It is very essential to note that these features are correlated and important for success of data mining process [4].

- **Constraint –Based perspective**. Data mining applications are always bounded by some rules as well as policies related to social, domain, financial, environmental aspects. Scalability, efficiency of algorithms, domain characteristrics, specific user's requirement,

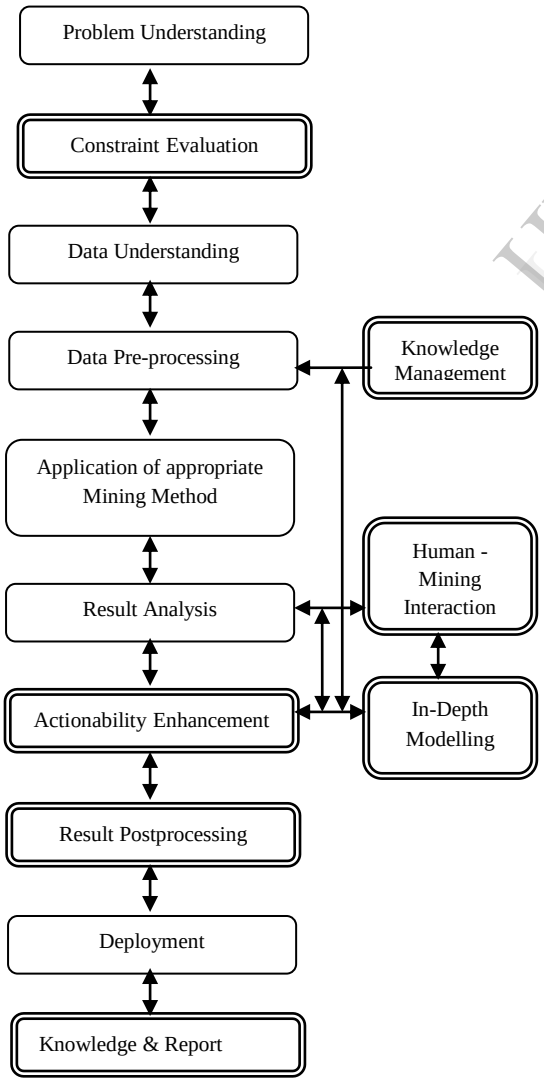

Figure 3. Data Mining & DDID-PD Framework

and types of deliverables are some examples of contraints need to be defined.

- **Integration of Domain knowledge with data to be mined.** In data mining process domain knowledge should be incorporated as background knowledge. One of the preferable ways of representation of domain knowledge is using Ontology. Domain knowledge should include concepts, beliefs, prefernces, and relations in business field. Presice domain knowledge can be represented in terms of relations (eg. is_a (A, B) →B). While vague knowledge need to be fuzzified & then mapped to precise terms & relations.
- ***Interaction between Human & mining system*.** Human being may need to play different roles during data mining process like user, business analyst, domain expert, result analyser etc. Human interaction with mining system is generally provided with explicitly designed forms to provide fine tunning of parameters, accessing knowledge base, visual interfacing etc. Interaction quality may depend upon user-friendliness of application, run-time capabilities, and representabilities.
- **Enhancing Knowledge Actionability.** Actionable patterns can be viewed as revised optimal version of the general patterns. They can be created by simply reducing, refining general patterns. But actionability of knowledge needs to be reviewed & revised iteratively in order to make patterns beneficial for business.
- ***Loop-Closed Iterative refinement.*** Actionable knowledge discovery process needs to have iterative feedback from various stages of KDD. Real world data mining applications also need to be highly iterative as constraint-analysis, feature-refinement, modelling, etc. can not be carried out just in one attempt. One should always keep in mind that real world data mining applications can not be deal with just an algorithm but needs an infrastructure, which will allow execution of various stages of KDD iteratively & interactively.
- **Creation of Reference model and questionnaire.** Reference models are designed to visualize how knowledge discovery will be carried out. Reference model should be domain specific. For example, in figure 4 a model for actionability enhancement has been shown. Questionnaire may be in form of a survey report, opinion poll or feedback form and is very necessary to build business requirement, plans, expected deliverables etc.

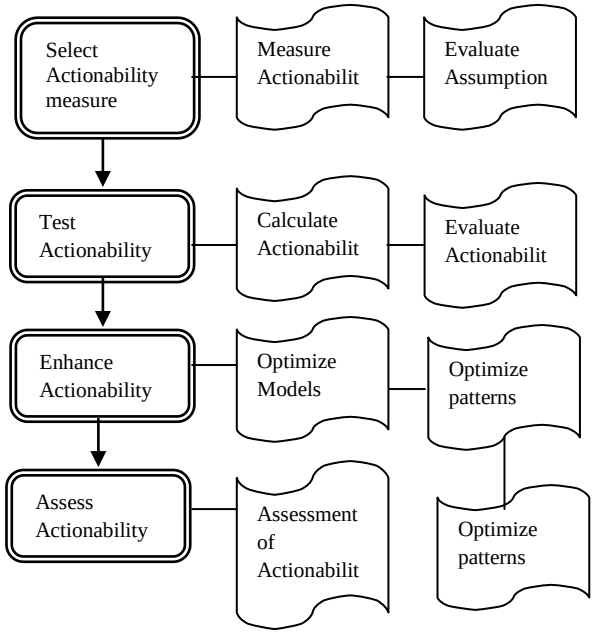

Figure 3. Model for actionability enhancement

5. Conclusion

Combinned mining is a general approach for designing of real life data mining applications. It includes designing frameworks for multifeature, multimethod, multisource approaches. Sample frameworks are discussed in this paper. More such domain specific frameworks can be designed. Also combined mining involve designing of various representation in which patterns can be delivered. For example pair pattern, incremental pair pattern, cluster pattern, incremental cluster pattern are elaborated. There is further scope to design new combined mining approaches, pattern merging methods, representation of deliverable patterns.

New era data mining applications don't want large number of rules, but reduced number of rules with decision making knowledge. DDID-PD framework provides a paradigm shift from data-driven pattern mining to domain driven actionable pattern discovery. Using this framework various domain specific models for telecom froud detection, trading evidence discovery, stock-market analysis, design of marketing policies can be built.

5. References

[1] Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE Transactions on systems, man and cybernetics, VOL. 41, No. 3, June 2011

[2]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.

[3]Cao et al., *Domain Driven Data Mining*, DOI 10.1007/978-1-4419-5737-5_1, Springer Science + Business Media, LLC 2010

[4] Cao L and et al. Domain-driven data mining: a practical methodology, *Int. J. of Data Warehousing and Mining*, 2(4): 49-65, 2006.