Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 6, Number 1 (2014), pp. 35-45

http://www.irphouse.com

Incorporation of Domain Knowledge for Post Mining of Combined Patterns

Mrs. Suvarna R. Bhagwat[1] and Prof. H. A. Hingoliwala[2]

[1] PG Student, JSCOE, University of Pune, Pune, Maharashtra, India
suvarnarbhagwat@gmail.com
[2] Faculty, JSCOE, University of Pune, Pune, Maharashtra, India

Abstract

Over last 10-15 years data mining practices have briskly employed in various fields. Input data for such miscellaneous applications not only varies but is also complex i.e. huge as well as diverse in nature. Many data mining applications are restricted to mine only a specific type of data. Moreover, these applications use solitary technique of data mining to discover knowledge.

Combined mining embraces three different framework viz. multi-source combined mining, multi-method combined mining, and multi-feature combined mining. The outcome of combined mining process is termed as combined patterns, which are indisputably more informative than simple association rules. The combined patterns take various forms as atomic patterns, pair patterns, cluster patterns, incremental pair patterns and incremental cluster patterns.

Many times data miner may perhaps not be satisfied with the outcomes of mining. His demand is to incorporate his expertise, demand or predefined goals into process of mining. This is nothing but the domain knowledge, which plays vital role in data mining applications. Out of various ways of representing domain knowledge, ontology is effective one. It not only helps to generalize the attributes in datasets but also provide a tactic to categorize them according to user's perspective. The combined patterns can be converted into more practicable information if one processes them with the help of domain knowledge.

Keywords: Combined Mining, Combined patterns, Ontology, Generalization

Introduction

Data mining, the extraction of hidden predictive information from large databases, is a

dominant technology. Data mining tools predict future trends and behaviors, allowing businesses to make practical, knowledge-driven decisions. There are various data mining techniques like classification, prediction, associations, clustering which help to represent the important data in various forms.

Association rules mining technique is widely used in many data mining application. Association rule mining always find intricacy in handling multiple datasets having numerous features. General approach for handling such huge datasets is to perform joining operation. But it is also not feasible in all situations. Also many times the number of rules generated does not reflect in general knowledge from all datasets.

An efficient approach for discovering association rules from huge, multiple sources, Combined mining has been proposed in [1]. The discovered patterns from combined mining are named as 'Combined Patterns'. They reflect acquaintance from important attributes of all considered datasets. Combined mining can be carried out in three approaches. Multi-source combined mining provide solution to handle involvement of multiple huge data sets during the process of data mining. Multi-feature combined mining assist to knob important features from multiple sources. Multi-method combined mining insist on use of multiple data mining techniques. This helps to incorporate pros and cons of different techniques.

This paper stand on multi-feature combined mining approach defined in [1]. Also we are representing an approach for incorporation of domain knowledge within the combined patterns. This make possible to generalize the knowledge represented by combined patterns. The base of this paper is work, which is carried out in following modules.

- **Use of Multiple sources:** Initially user is able to select required sources from the available ones.
- **Use of Significant attributes:** Not all but only significant attributes from each of the selected datset are further ahead used for discovering various patterns.
- **Pattern discovery:** Not only atomic but pair patterns, incremental pair patterns, cluster patterns as well as incremental cluster patterns are generated using the significant attributes. All these patterns are named as combined patterns. For this discovery process unconventional interestingness measures like I_{rule}, I_{pair}, $I_{cluster}$ are applied [1].
- **Incorporation of domain knowledge:** Combined patterns reflect significant attributes from various datasets. But now and then data miner possibly will wish insert his expertise into the results. In this module data miner's expertise is represented in form of ontology. Combined patterns are then refined to reflect the domain knowledge.

Such a refinement makes the combined patterns equipped to use in decision making process and thus more feasible.

Literature Survey

Knowledge discovery in database (KDD) is an automated process to identify useful

and comprehensible patterns from large datasets. Data preprocessing, data mining and post processing are major steps in process of KDD. Data mining is most decisive process in knowledge discovery. Out of several data mining techniques, association rule mining aid to discover interesting patterns from input dataset.

Association rule take form like X-->Y. Every association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is set up in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is used to indicate of how frequently the items appear in the database. Confidence specify the number of times the if/then statements have been found to be true. Briefly, Support = no. times X appears in dataset.

Confidence = $P(X \cap Y)/P(X)$

Research on Unconventional interestingness measures

Support and confidence are very successful interestingness measures vigorously used by data miners worldwide. The main problem faced by association rule mining is generation of huge number of rules along with the preeminent ones. Some unconventional interestingness measures like Lift, I_{rule}, I_{pair}, $I_{cluster}$ are brought into attention by Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang [1]. These measures are derived from support and confidence by mathematical formulas. They help to trim down the number of rules generated by association rules mining.

Research on combined patterns

Discovery of more informative or sensible association rules is done generally through post analysis, joining multiple data sets or by combining multiple data mining methods. All these approaches may not be feasible in all circumstances. Combined patterns are represented as A1, A_2….. A_i, B_1, B_2, $B_j \rightarrow$ S where A_i and B_j are item sets from different datasets and S is target item or item set[1]. The authors have proposed four novel patterns under multi-feature combined mining approach, in which traditional association rules can be transformed using various interestingness measures. These are pair patterns, cluster patterns, incremental pair patterns and incremental cluster patterns.

Research on Ontology

Combined patterns reflect knowledge from different input data sets. They are obviously more informative as compared to association rules. But in some circumstances, data miner may not expect the detailed or low level knowledge. For example a combined pattern

{AGE_YOUNG, HAVING_BIKE, SALARY_HIGH}→{BUY_CAR}

provides probability of purchasing a car by a young, salaried bike owner. Similarly,

{AGE_YOUNG, HAVING_BIKE, SALARY_HIGH}→{BUY_BIKE}

provides probability of purchasing a second bike by a young, salaried bike owner. But data miner may wish to get answer of "What is probability of purchasing a vehicle (Bike or Car) by a young, salaried person who already own a vehicle (Bike or Car again)?" That means data miner possibly will need to generalize the knowledge.

Domain knowledge can be represented in form of rules, if-then statements, is-a relationship, decision trees etc. Some knowledge specification languages like General Impressions (GI), Reasonably Precise Concepts (RPC), and Precise Knowledge (PK) are used to build "is-a organization" of database attributes. This type of organization is called as "Item categorization". It helps to represent knowledge in form of hierarchy. Using ontology concepts one can extend these languages [2]. In this paper ontology has been used to generalize simple association rules through an interactive framework 'ARIPSO' in.

Combined patterns

Authors have designed three assorted frameworks which utilize the already existing data mining tools to get more informative patterns [1]. The outcomes of combined mining are named as combined patterns. Various types of combined patterns are discussed in this section. Consider following example data set for further discussion.

Table 1: Example Dataset

Record	A	B	C	D	E	F
1	1	1	1	0	0	0
2	1	0	1	0	0	0
3	1	0	0	1	0	0
4	0	1	0	0	1	1

Atomic patterns

Some atomic patterns along with their traditional interestingness measures derived from above dataset are revealed in Table 2.

Table 2: Citation of atomic patterns

Atomic patterns	Support Prob(X∩Y)	Confidence Prob(X∩Y)/Prob(X)	Lift Prob(X∩Y) / Prob (X) * Prob(Y)
A→B	1/4	(1/4) / (3/4)= 1/3	(1/4) / (3/4) * (2/4)= 2/3
A→C	1/2	(1/2) / (3/4)= 2/3	(1/2) / (3/4) * (2/4)= 4/3
A→D	1/4	(1/4) / (3/4)= 1/3	(1/4) / (3/4) * (1/4)= 4/3
B→C	1/4	(1/4) / (2/4)= 1/2	(1/4) / (2/4) * (2/4)= 1
C→B	1/4	(1/4) / (2/4)= 1/2	(1/4) / (2/4) * (2/4)= 1
E→F	1/4	(1/4) / (1/4)= 1	(1/4) / (1/4) * (1/4)= 4

Pair patterns

Pair patterns can be formed by combining two atomic patterns. They are of form {A1 --> B1, A2 -->B2} where A1 and A2 are same but B1 and B2 are different or vice versa. Measure I_{pair}, is used to measure the interestingness of pair pattern.

In case of pair patterns three new interestingness measures are considered, called contribution, I_{rule} and I_{pair}. They are defined as follows.

Contribution (X, Y→Z) = lift(X, Y→Z) / lift(X-→Z)

= confidence (X, Y→Z) / lift(X→Z)

I_{rule} = I $_{rule}$ (X, Y→Z) = Lift (X, Y→Z) / Lift (X→Z) * Lift (Y→Z)

I_{pair} = |conf(X-→Z_1) – conf(Y→Z_2) |

If $Z_1 = Z_2$

= square-root(confidence(X→Z_1) *conf(Y→Z_2)

…. If $Z_1 \neq Z_2$

= 0….Otherwise

Table 3 shows pair patterns generated from atomic patterns in Table 2.

Table 3: Citation of Pair patterns

Pair Pattern	Prob(X∩Y)	Prob(X)* Prob(Y)	Lift (XY→Z)	Lift (X→Z)	Lift (Y→Z)	I $_{rule}$	I_{pair}
(A, C → B)	1/4	2/4 *2/4	1	2/3	1	1.5	1/6
(A, B→ C)	1/4	1/4* 2/4	2	4/3	1	1.5	1/6
(B, E →F)	1/4	1/4* 1/4	4	2	4	0.5	1/2
(B, F) →E)	1/4	1/4* 1/4	4	2	4	0.5	1/2

Incremental pair patterns

Some patterns can take form of extension of other patterns. For example, {A1, B1 □C1} can be thought of an extension of {A1 □ C1}. Such patterns are combined to form incremental pair patterns.

The conditional 'Piatetsky–Shapiro's ratio', Cps has been defined for calculating interestingness 'Incremental pair patterns' of as follows.

$Cps(B \rightarrow C|A)$ = *Probability* $(B \rightarrow C / A) - Probability\ (B / A) \times Probability\ (C/A)$

In above discussed example {(A□C), (A, B□C)} can be deliberated as incremental pair pattern.

Cluster patterns

Cluster patterns, are formed by organizing many similar or related atomic or pair patterns together. They take the form, {A1 □ B1, A2 □ B2, …, AN □ BN}. Measure $I_{cluster}$ defines how interested is the cluster of patterns. Such cluster of patterns is more informative than their integral patterns. The maximum interestingness of any pair pattern within a cluster is considered as $I_{cluster}$.

From above example dataset {(A□C), (B□C), (D□C)} is a cluster of patterns, with (A, B□C) having maximum I_{pair} value of 1/6.

Incremental cluster patterns

Many patterns which are extension of one another can be grouped together to form Incremental cluster patterns, as {(A1 □ Z1), (A1, B1 □ Z1), (A1, B1, C1 □ Z1)...}. Note that 'impact' a new interestingness measure has been demarcated to calculate the impact of incremented portion of rule on existing pattern. It is defined as follows.

If contribution of a pattern, P > = 1 then it's impact = contribution (P) – 1. Else impact (P) = [1 / contribution (P)] -1.

Knowledge representation using ontology

Is-a hierarchy is used to represent generalized knowledge. Using ontology we can extend the hierarchy to incorporate user's perspective about the circumstances. Three concepts are used to construct ontology viz. Leaf concepts, generalized concepts and restricted concept[2].

Leaf concepts

A leaf concept is defined such that, each leaf concept is linked to one item in the database. Figure 3 shows extract of ontology for a shopping database. E.g. Full shirt, half cargo pants etc.

Generalized Concepts

Generalized concepts are defined such that the concepts subsume other concepts in the ontology. A generalized concept is associated to the database through its subsumed concepts. E.g. Formal, Casual.

Restricted concepts

Restriction concepts are described knowledge of domain expert and depend on the user individually. E.g. Summer wear, winter wear etc.

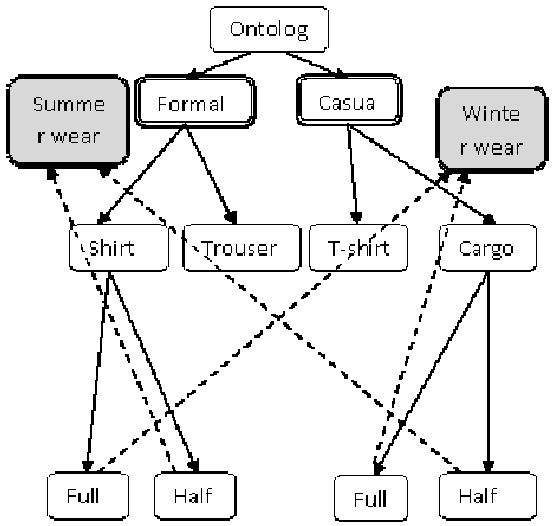

Figure1: Example ontology

Experimental Setup

As already discussed the combined mining methodology provides three diverse frameworks viz. multi-source combined mining, multi-feature combined mining and multi-method combined mining. We have taken efforts on multi source and multi feature combined mining frameworks. These frameworks can be applied on multiple datasets data sets on ad hoc basis. This process will discover "combined patterns" form input data sets. As shown in figure 1.

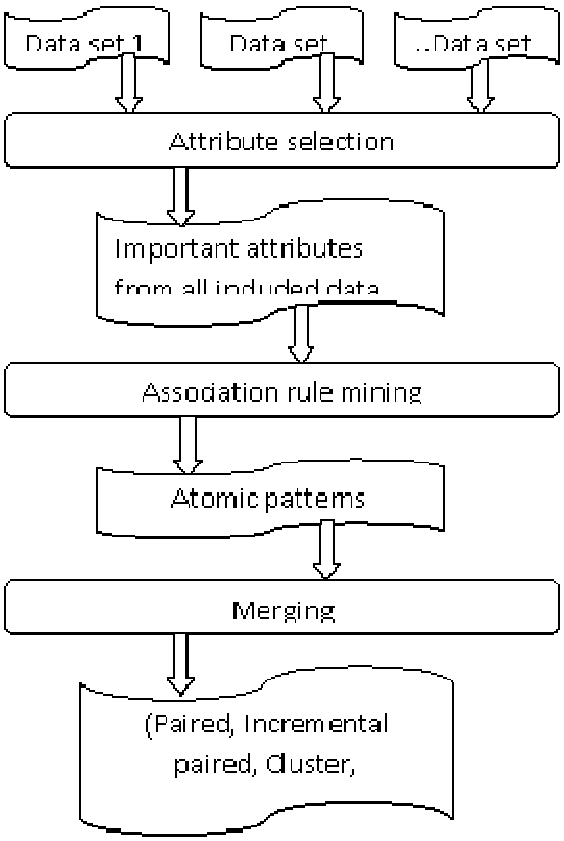

Figure2: Experimental setup for multi-source and multi-feature combined mining.

The "combined patterns" are surely more informative than simple association rules. But many times they represent low level or detailed knowledge. In certain circumstances data miner requires generalized facts. To reinforce the incorporation of domain knowledge in post-mining process it is represented with the help of ontology. The post processing of combined patterns done in two steps.

Formation of domain knowledge base:

Domain knowledge offers a general view over user knowledge in database domain, and user expectations express the prior user knowledge over the discovered rules. In

this stride, the domain knowledge and goals are formalized using restricted concepts in the ontology.

Application of domain knowledge for post mining process:
The next pace is to improve the efficacy of combined patterns by generalization. The combined patterns discovered by combined mining process can be simplified with indulgence of domain knowledge (represented using ontology).

From above discussion it clear that user plays important role in mining as well as in post mining process. User has to decide on ad hoc basis about the framework to be used for mining. On the other side he also contributes to build domain knowledge. After discovery of combined patterns domain knowledge will be incorporated for their post processing. The overall interaction of user in whole process is depicted in figure 3.

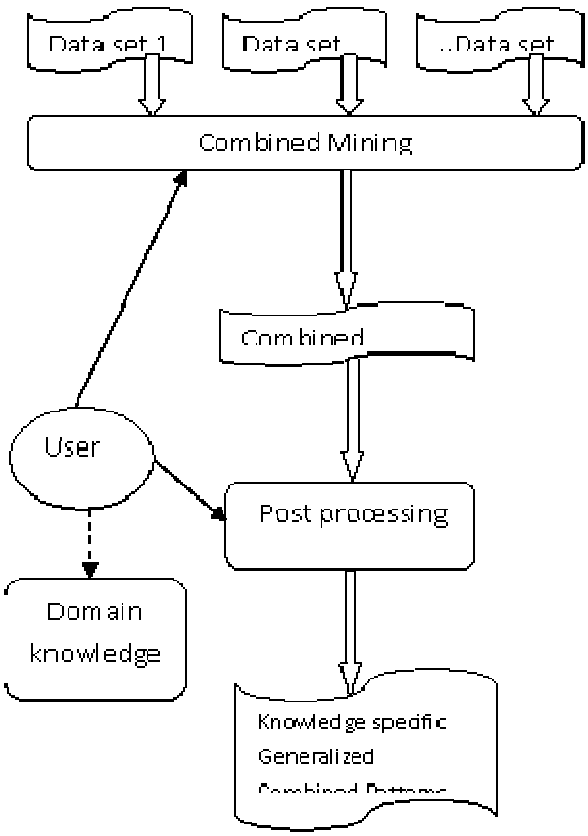

Figure3: Interaction of user in mining, domain knowledge building and post mining.

Illustration of data set
The implementation of above mentioned experimental set up is done using following data sets. The datasets are manually created. The combined mining will be able to answer the questions like-

- What is probability purchasing a car by a young, high salaried male already possessing a bike?
- How many married females having a Fridge will buy a Washing machine?
- How does marital status affect purchasing of a kind of vehicle?
- What is probability of purchasing a desktop by unmarried male who is own a laptop?

Table 4: Data set1-Personal Information

Attributes	Description
Gender	Male: 0 Female: 1
Marital Status	Unmarried: 0 Married: 1
Age_ young	If age < 25 : 1 else 0
Age_ middle	If 25 < age <50 : 1 else 0
Age_ old	If age > 50 : 1 else 0
Income_ class1	If income < 20000 : 1 else 0
Income_ class2	If 20000 < income < 40000 : 1 else 0
Income_ Class3	If 40000 < income < 70000: 1 else 0
Income_ class4	If income >70000 : 1 else 0

Table 5: Data set2-Prior Possessions

Attributes	Description
Bike	1 if possesses, 0 if not
Car	1 if possesses, 0 if not
Mobile	1 if possesses, 0 if not
Laptop	1 if possesses, 0 if not
Fridge	1 if possesses, 0 if not
Food Processor	1 if possesses, 0 if not
Washing Machine	1 if possesses, 0 if not

Table 6: Data set3-Recent Purchase

Attributes	Description
Bike	1 if purchased recently, 0 if not
Car	1 if purchased recently, 0 if not
Mobile	1 if purchased recently, 0 if not
Laptop	1 if purchased recently, 0 if not
Fridge	1 if purchased recently, 0 if not
Food Processor	1 if purchased recently, 0 if not
Washing Machine	1 if purchased recently, 0 if not

On top of the above dataset user's perspective is characterized by ontology extract of ontology is described in figure 4.

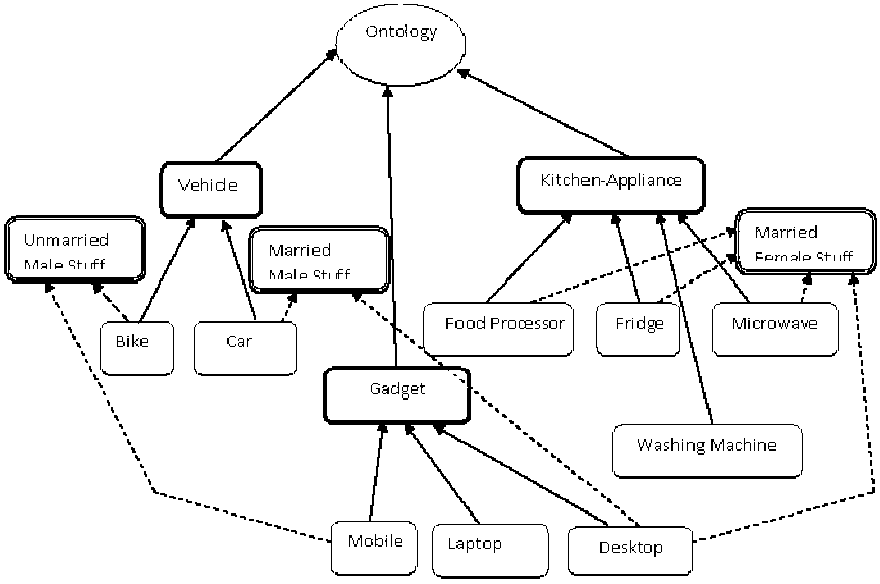

Figure 4: Ontology representation

Here in above ontology leaf concepts are bike, car, food processor, fridge, microwave, washing machine, mobile laptop and desktop. They are naturally categorized into vehicle, gadgets, kitchen appliances using simple is-a relationship, named as generalized concepts in ontology..

But after a while data vender will come to know that mobiles are more purchased by a unmarried male as compared to married male or females. This type of domain knowledge is represented by restricted concept 'Unmarried male stuff'. Similarly car and desktop are purchased more by married male.

Conclusions and Performance analysis

Association rule mining has been vigorously used and researched by worldwide data miners. The outcome represents interesting patterns among the attributes in data set. But massive numbers of rules are generated in most of the circumstances. Also many times rules represent knowledge from a single dataset. If user want to incorporate many data sets then expensive table joining operations are necessary. By use of multi-source combined mining we are able to avoid table joining. The multi feature combined mining helps to concentrate on most significant attributes from available data set, which on the other hand brings out So only sensible rules. Such rules cannot be directly discovered by any primitive algorithm. In short the outcomes of multi-source and multi-feature combined mining integrate important facts from multiple data sets and also reduce number of rules generated.

Such combined patterns are definitely more practical as compared to simple association rules also they are to a large extent less in number. But in certain circumstances where data miner's perspective is also equally important they cannot be used directly. Representation of domain knowledge using ontology is exceptionally much effective and easy to implement. Ontology helps out to generalize the items in data set. When data miner does not need detailed or thorough acquaintance of the data set, combined patterns can be generalized with the assistance of ontology. Generalized combined rules imply a unusual knowledge from various datasets.

References

[1] L. Cao, Y. Zhao, and C. Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE Transactions on Systems, Man, and Cybernatics—PART B: CYBERNETICS, VOL. 41, NO. 3, JUNE 2011, 699.

[2] Claudia Marinica and Fabrice Guillet, "Knowledge- Based Interactive Postmining of association rules using ontologies", IEEE Transactions on knowledge and data engineering, VOL. 22, NO. 6, JUNE 2010.

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

[4] 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.

[5] Cao L and Zhang C. Domain-Driven Data Mining: A Practical Methodology, International Journal of Data Warehousing and Mining, 2(4):49-65, 2006.

[6] B. Baesens, S. Viaene, and J. Vanthienen, "Post-Processing of Association Rules, " Proc. Workshop Post-Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics with Sixth ACM SIGKDD, pp. 20-23, 2000.

[7] M. Uschold and M. Grul ninger, "Ontologies: Principles, Methods, and Applications, " Knowledge Eng. Rev., vol. 11, pp. 93-155, 1996.

[8 Local Mining of Association Rules with Rule Schemas Andrei Olaru Claudia Marinica Fabrice Guillet LINA, Ecole Polytechnique de l'Universite de Nantes rue Christian Pauc BP 50609 44306 Nantes Cedex 3 E-mail: cs@andreiolaru.ro, fclaudia.marinica, fabrice.guilletg@univ-nantes.fr