Prashasti Kanikar, Dr. Ketan Shah / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.1295-1300

# Extracting Actionable Association Rules from Multiple Datasets

**Prashasti Kanikar*, Dr. Ketan Shah****
*Department of Computer Engineering,MPSTME,NMIMS (Deemed to be University) , Mumbai- 400056,INDIA
*Department of Information Technology,MPSTME,NMIMS (Deemed to be University) , Mumbai- 400056,INDIA

## ABSTRACT
**Applying traditional association rule mining approaches on multiple datasets generally results in generating large number of redundant association rules. Finding actionable association rules from this is very tedious and sometimes infeasible also. In the paper , we apply the Combined Mining approach using *I*rule parameter for generating actionable association rules. The approach is tested on a survey data set that consists of multiple related data items collected independently. The results indicate that combined mining approach helps in generating actionable association rules.**

**Keywords**
Association Rule Mining, Data Mining, Knowledge Discovery in Databases, Pattern Mining.

## 1. INTRODUCTION
Knowledge discovery in database (KDD) is an active area of research that resolves the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. In other words, the origin of KDD, many researchers realized the need from 'data' to 'knowledge' for the business decision-making, such as [5-7]. Recently, more efforts have shifted from 'valid' and 'understandable' knowledge to actionable knowledge especially for real world data mining applications. In simple terms, a pattern is actionable if the user can act upon it to her advantage. Furthermore, actionable patterns can not only afford important grounds to business decision-makers for performing appropriate actions, but also deliver, expected outcomes to business.

Association rule mining is a main method to produce patterns. However, as large numbers of association rules are often produced by association mining algorithm, sometimes it can be very difficult for decision makers to not only understand such rules, but also find them a useful source of knowledge to apply to the business processes. In other words, association rules can only provide limited knowledge for potential actions. Therefore, there is a strong and challenging need to mine for more informative and comprehensive knowledge for decision-making in the real world. A
comprehensive and general approach for discovering informative knowledge in complex data is suggested.

It is challenging to mine for comprehensive and informative knowledge in such complex data suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, 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. The solution to these problems is "combined mining".

## 2. RELATED WORK

The notion of association rules was proposed 18years ago and is widely used today. Osmar R. Za¨ıane and Maria-Luiza Antonie have proposed strategies for classification rule pruning in the case of associative classifiers [3]. Yanchang Zhao, Huaifeng Zhang, Fernando Figueiredo, Longbing Cao, Chengqi Zhang have proposed a technique to discover combined rules on multiple databases and applied to debt recovery in the social security domain[10].

In the area of pattern mining also, variety of approaches are suggested. Zhiwen Yu, Xing Wang ,Hau-San Wong and Zhongkai Deng [4] have used normalized cut approach which finds the local patterns from six challenge datasets designed By author. Shigeaki Sakurai , Youichi Kitahara, and Ryohei Orihara [5] proposed the sequential interestingness as a new evaluation criterion that evaluates a sequential pattern corresponding to the interests of analysts. Unil Yun, and John J. Leggett [6] proposed an algorithm (WSpan) that generates fewer but important weighted sequential patterns in large databases, particularly dense databases with a low minimum support, by adjusting a weight range.

## 3. COMBINED MINING
Usually, there are a lot of tasks of data mining, such as classification, clustering and rule mining. Among of these tasks, the mining of patterns for discerning relationships between data items in large databases is a well studied technique. In order to introduce the key research, more details of association rule mining, combined association rule mining and association rule mining with composite items will be presented in this section.
**Association rule mining**, a widely used data mining technique, is used to reveal the nature and frequency of relationships or associations between entities. Support and confidence, are the two major indices, which have useful applications to evaluate the rules. For instance, consider rule *X*: if *E* then *F*. Suppose that *X* has 60% confidence

and 40% support. It expresses that 40% of records contain *E* and *F*. In fact, this means that in 40% of total records, rule *X* is valid. Additionally, it expresses that 60% of records that contain *E*, contain *F* as well. However, conventional association rules, as discussed above, can only provide limited knowledge for potential actions.

Strictly speaking, traditional association rule mining can only generate simple rules. However, the simple rules are often not useful, understandable and interesting from a business perspective. Thus, Zhao et al. [1] proposed combined association rules mining, which generated through further extraction of the learned rules. In other words, to present associations in an effective way, and in order to discover actionable knowledge from resultant association rules, a novel idea of combined patterns is proposed.

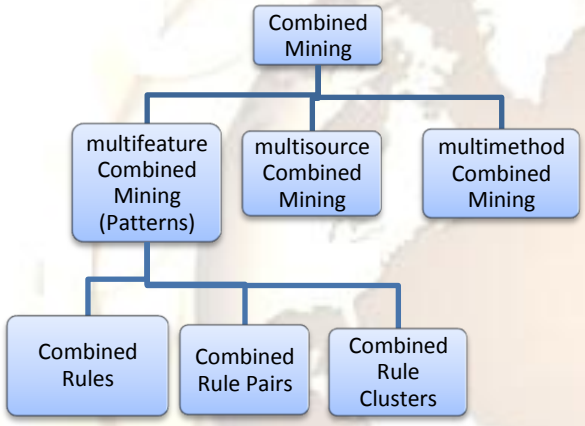


Fig 1: Classification of Combined Mining Approach ([1] and [2])

In **multifeature combined pattern 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.

**Multimethod combined mining** is another approach to discover more informative knowledge in complex data. The focus of multimethod combined mining is on combining multiple data mining algorithms as needed in order to generate more informative knowledge. In fact, the combination of multiple data mining methods has been recognized as an essential and effective strategy in dealing with complex applications.

**A combined association rule** is composed of multiple heterogeneous itemsets from different datasets. A combined rule pair is composed of two contrasting rules and **combined rule clusters** are built from combined association rules. The proposed combined patterns provide more interesting knowledge and more actionable results than traditional association rules [1].

Compared with simple association rules, combined association rules have more chances to take actions for decision makers in real data mining applications. The following example explains the point.

### 3.1 Advantages of Combined Approach

Advantages of combined mining in discovering informative knowledge in complex data, compared to a single use of existing methods.

**1)** Flexible frameworks for combining multifeatures, multisources and multimethods covering various needs in mining complex data, which are customizable for specific cases. With combined mining, the advantage of specific algorithms can be well taken in handling particular tasks.

**2)** Effective in discovering patterns with constituents from multiple heterogeneous sources and a large scale of real life data, which can provide patterns reflecting a full picture rather than a single line of business.

### 3.2 *I*rule Calculation

*I*rule indicates whether the contribution of *U* (or *V)* to the occurrence of *T* increases with *V* (or *U*) as a precondition. Therefore, "*I*rule < 1" suggests that $U \cap V \rightarrow T$ is less interesting than $U \rightarrow T$ and $V \rightarrow T$. The value of *I*rule falls in [0,+∞). When *I*rule > 1, the higher *I*rule is, the more interesting the rule is. Therefore, our new measures are more useful than the traditional confidence and lift [1].

$$I\text{rule}\,(U \wedge V \rightarrow T) = \frac{Lift\,(U \wedge V \rightarrow T)}{Lift\,(U \rightarrow T) * Lift\,(V \rightarrow T)}$$

## 4. STEPS FOR COMBINED MINING

In following diagram, the steps for implementing combined association rule mining are given.

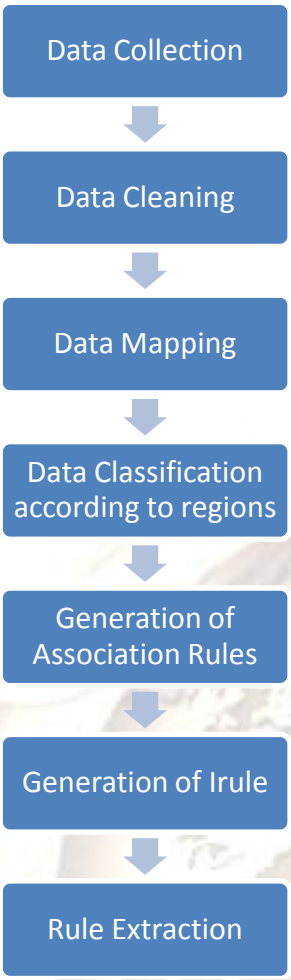


Fig. 2: Steps for Combined Mining

## 5. DATA SET DESCRIPTION

The data set considered here is taken from a survey report to find out quality rules in order to make decisions to make the travelling convenient for the people of that particular region. The data set consists of 1794 records with following attributes and the answers of Questions asked to people of that area.

Table I
Data set considered

| ATTRIBUTE | DESCRIPTION |
|---|---|
| **Gender** | Male or Female |
| **Age** | Given in years |
| **Marital** | Marital Status of respondent |
| **Level of Education** | Highest Level of Education Achieved |
| **Gross Income** | Gross Income of respondent |
| **Region** | Region where respondent lives |
| **Question 1** | Method of Transport to do Main Shop |
| **Question 2** | Time Taken to Travel to do Main Shop |
| **Question 3** | Ease of Travel to Main Shop |

## 6. EXPERIMENTATION PERFORMED

### 6.1 Data Mapping

The textual data is mapped to numeric values in order to make the computations smoother.

**Gender**

- 1- female
- 2- male

**Marital status**

- 3- single
- 4- married
- 5- divorced
- 6- widowed
- 7- separated

**Gross income**

- 31- Less than £2600
- 32- £2600 to less than £5200
- 33- £5200 to less than £10400
- 34- £10400 to less than £15600
- 35- £15600 to less than £20800
- 36- £20800 to less than £28600
- 37- £28600 to less than £36400
- 38- £36400 or more
- 41- Refused
- 42- Unknown

**Method of travel**

- 51- Car
- 52 -Public Transport
- 53- Don't Shop
- 54- On Foot
- 55- Other
- 56- Bicycle
- 57- Motorbike/Moped

**Time taken**

- 61- 5 mins or less
- 62- 6-10 mins
- 63- 11-20 mins
- 64 -21-30 mins
- 65 -31-45 mins
- 66- 45 + mins
- 70- N/A

**Ease**

- 71- Very Easy
- 72 -Fairly Easy
- 73- Fairly Difficult

- 74- Very Difficult
- 80- N/A

**Region**

- 101- The North
- 102 -South West
- 103 -Midlands/East Anglia
- 104 -London
- 105 -South East
- 106 -Wales
- 107 -Scotland

### 6.2 Data Classification

The data is classified on the basis of regions. The total number of record are 1794. Here, the data consists of seven regions so it is classified into seven classes. Data from each class is processed further for generating association rules.

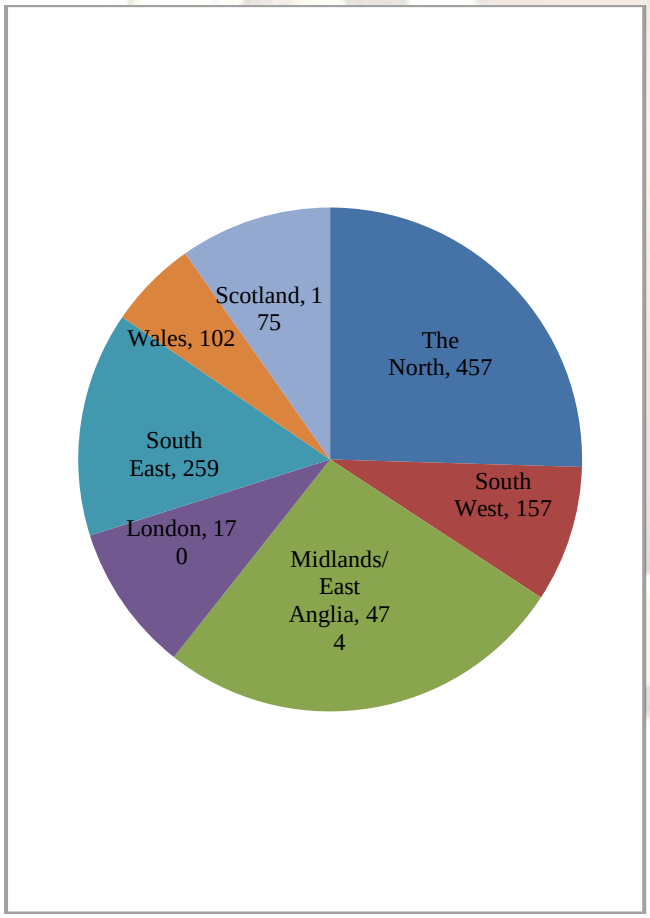


Fig.3: Region based classification of 1794 records

### 6.3 Generation and Extraction of Association Rules

First the association rules are generated using apriori algorithm. After that Lift of the rule is computed. Then Interestingness measure *I*rule is computed for all the rules. Table II shows the rules with *I*rule value greater that 1.

## 7. RESULTS AND DISCUSSION

In Table II the association rules are given in first four columns in antecedents -> consequents format. In **Lift** column, the lifts of rule are computed using the formula given in section 4.1. **LDenomLeft** and **LDenomRight** are the lifts of left and right parts of the rule. For example in rule {female,Very easy}->{6-10 minutes} the **LDenomLeft** is the lift of rule {female}->{6-10 minutes}and **LDenomRight** is the lift of rule { Very easy }->{6-10 minutes}. **DLift** is the product of **LDenomLeft** and **LDenomRight.** Finally ***I*rule** is

Table II
Rules Extracted for region 1(The North) when support =10% and confidence=10%

| Antecedent 1 | Antecedent 2 | -> | Consequent | Lift | LDenomLeft | LDenomRight | DLift | Irule |
|---|---|---|---|---|---|---|---|---|
| {female | Very easy} | -> | {6-10 minutes} | 1.44 | 1.04 | 1.2 | 1.248 | 1.15385 |
| {car | fairly easy} | -> | {11-20 minutes} | 1.35 | 0.9 | 1.44 | 1.296 | 1.04167 |
| {female | 6-10 minutes} | -> | {Very easy} | 1.32 | 0.9 | 1.2 | 1.08 | 1.22222 |
| {female | car} | -> | {6-10 minutes} | 1.24 | 1.04 | 1.15 | 1.196 | 1.03679 |
| {car | fairly easy} | -> | {6-10 minutes} | 1.19 | 1.15 | 1.03 | 1.185 | 1.00464 |
| {male | car} | -> | {Very easy} | 1.18 | 1.06 | 1.1 | 1.166 | 1.01201 |
| {£5200 to less than £10400 | car} | -> | {female} | 1.18 | 1.16 | 0.9 | 1.044 | 1.13027 |
| {female | 6-10 minutes} | -> | {car} | 1.17 | 0.9 | 1.15 | 1.035 | 1.13043 |
| {6-10 minutes | fairly easy} | -> | {car} | 1.16 | 1.15 | 1 | 1.15 | 1.0087 |
| {male | Very easy} | -> | {car} | 1.14 | 1.02 | 1.1 | 1.122 | 1.01604 |
| {6-10 minutes | Very easy} | -> | {female} | 1.14 | 1.04 | 0.9 | 0.936 | 1.21795 |
| {car | Very easy} | -> | {male} | 1.1 | 1.02 | 1.06 | 1.081 | 1.01739 |
| {male | fairly easy} | -> | {car} | 1.09 | 1.02 | 1 | 1.02 | 1.06863 |
| {female | car} | -> | {£5200 to less than £10400 } | 1.09 | 1.16 | 0.9 | 1.044 | 1.04406 |
| {female | Very easy} | -> | {car} | 1.06 | 0.9 | 1.1 | 0.99 | 1.07071 |
| {car | 6-10 minutes} | -> | {female} | 1.06 | 0.9 | 1.04 | 0.936 | 1.13248 |
| {car | fairly easy} | -> | {female} | 1.04 | 0.9 | 1.11 | 0.999 | 1.04104 |
| {car | 6-10 minutes} | -> | {fairly easy} | 1.04 | 1 | 1.03 | 1.03 | 1.00971 |
| {female | car} | -> | {Very easy} | 1.03 | 0.9 | 1.1 | 0.99 | 1.0404 |

computed by dividing **Lift** by **DLift.** Here ,only the rules with Irule value greater than 1 are shown because if *I*rule is less than 1 then it means that individual rules are much stronger than combined rules.

Using the apriori algorithm , a large number of rules are generated but we can not apply Combined Mining approach on the rules containing a single antecedent . Fig. 6 gives details of number of rules generated using apriori algorithm and number of rules extracted for North region using the *I*rule parameter. Taking support 10% and confidence 10% , 102 association rules are generated and Combined Mining is applied on 33 rules. Out of these, 14 rules(42%) are rejected. Taking support 20% and confidence 20% , 61 association rules are generated and Combined Mining is applied on 20 rules. Out of these, 11 rules(55%) are rejected.

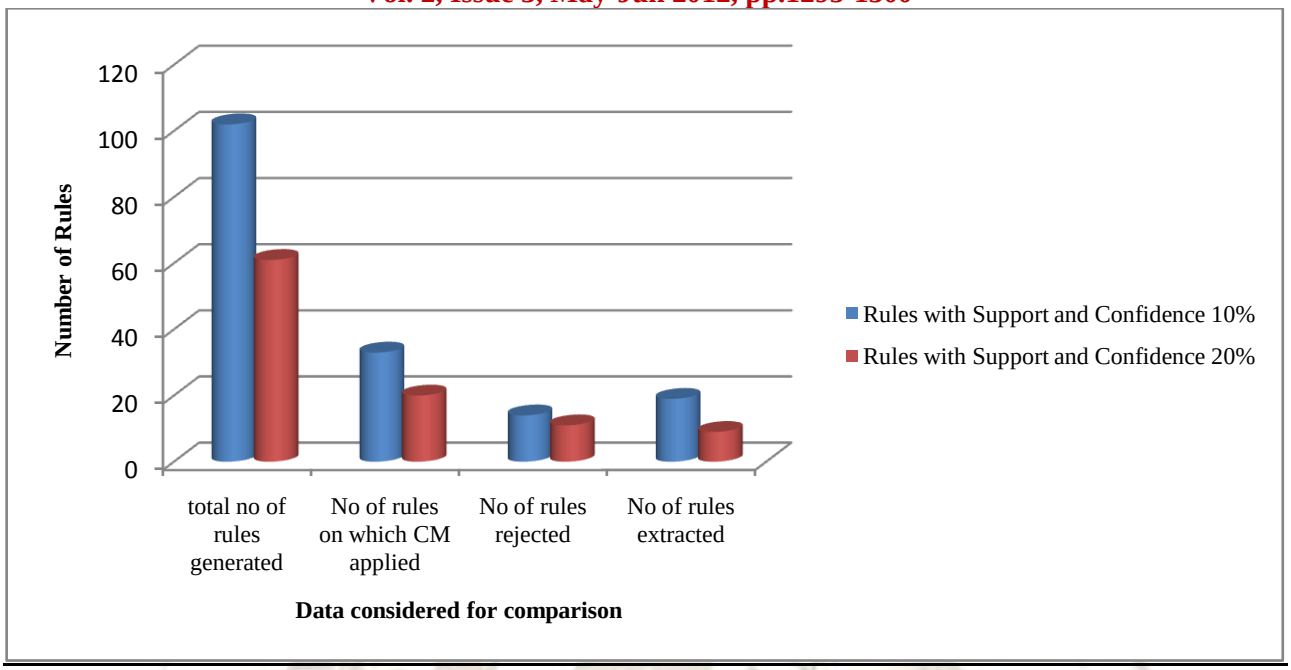


Fig .4: Results for The North region with varying Support and Confidence

## 8. CONCLUSION

The approach of Combined Mining using the interestingness measure *I*rule is presented in the paper. This approach allows us to filter out actionable rules from a larger set of rules generated by traditional algorithms.

## 9. REFERENCES

[1] 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.

[2] 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, JUNE 2011, pp. 699-712.

[3] Za¨ıane, O.R., Antonie, M.-L." On pruning and tuning rules for associative classifiers", KES 2005. LNCS (LNAI), vol. 3683, pp. 966–973.

[4] Yanchang Zhao,Huaifeng Zhang,Fernando Figueiredo,Longbing Cao,Chengqi Zhang," Mining for Combined Association Rules on Multiple Datasets", *2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007)*, August 12, 2007, San Jose, USA.

[5] H. Xu, G. L. Liu, J. Guo and Z. L. Lou, "Grain Price Impact Analysis System on Data Mining", JCIT: Journal of Convergence Information Technology, Vol. 6, No. 1, pp. 207 -211, 2011.

[6] W. S. Pan and P. W. Chen, "A Study on the Logistic Service Satisfaction for Internet Marketing Enterprise Using Data Mining Technology ", AISS: Advances in Information Sciences and Service Sciences, Vol. 3, No. 2, pp. 114-120, 2011.

[7] P. S. Wang, "Survey on Privacy Preserving Data Mining", JDCTA: Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, pp. 1 -7, 2010.

[8] Lingjuan Li, Min Zhang , "The Strategy of Mining Association Rule Based on Cloud Computing" , International Conference on Business Computing and Global Informatization (BCGIN) , 2011, pp. 475-478.

[9] Ashish Mangalampalli, Supervised by Vikram Pudi," Fuzzy Associative Rule-based Approach for Pattern Mining and Identification and Pattern-based Classification", *WWW 2011*, March 28–April 1, 2011, Hyderabad, India,pp. 379-383.

[10] T. Brijs, K. Vanhoof, G. Wets," Defining Interestingness for Association Rules", International Journal "Information Theories & Applications.