Ultra Scientist Vol. 26(3)B, 290-294 (2014).

Discovering Association Rules With Feature Selection Method Under Population Data

RAMESH PRASAD AHARWAL

Asstt. Prof., Department of Mathematics
Govt. P.G. College, Damoh, M.P. (India)

(Acceptance Date 14th November, 2014)

Abstract

This paper presents an application of ARM to discover precious association rules for survey data. In this paper we have used feature selection method with WEKA software. The aim of this paper presents the use of data mining approach in the social impact.

Key words: Data Mining, Association Rule, Feature selection, WEKA.

1. Introduction

Association Rule Mining is a technique used to discover relationships among a large set of variables in a data set. The concept of association rule mining firstly proposed by Agrawal, Imielinski, and Swami[1], Association Rule Mining related to the discovery of relationships among a large set of variables. It is given a database of records, each containing two or more variables and their respective values. Association rules come from market basket analysis and capture information such as "if customers buy book *X*, they also buy book *Y*". This can be written as $X \rightarrow Y$. Two measures characterize support and confidence an association rule is used. Data mining is the key step in the knowledge discovery process, and association rule mining is a very important research topic in the data mining field[1].

1.1 Basic Concept of Association Rule :

The Classical definition of association rules, as presented in[2] and [4], is: Let $\{t_1, ..., tn\}$ be a set of transactions, and let I be a set of items, $I = \{i_1, i_2 \ldots i_m\}$. Let D, the task-relevant data, be a set of transactions where each transaction T is a set of items such that $T \in I$. Let X be a set of items. A transaction T is said to contain X if and only if $X \in T$. An association rule is an implication of the form $X \in Y$, where $X, Y \in I$, and X and Y are disjoint item sets, i.e. $X \cap Y = \Phi$. Below we present

the algorithm used to mine association rules. Detail description of Association rule mining can be found in data mining literatures and (Han & Kamber[4]).

1.2 Apriori Algoritm :

A number of efficient association rule mining algorithms have been proposed in the last few years. Among these, the Apriori algorithm[4] has been very influential. Since its inception, many scholars have improved and optimized the Apriori algorithm and have presented new Apriori-like algorithms[3]. The Apriori-like algorithms adopt an iterative method to discover frequent *itemsets*. The Apriori algorithm is so named because it is based on the fact that it uses prior knowledge of frequent *itemset* properties known as the Apriori property. These terms are defined as follows. An *itemset* is any subset of all the items in the database of transactions. Below we present the pseudo code for the Apriori algorithm, as shown in[4].

1.4 Metrics Used in Association Rules:

Two metrics are usually used to measure the reliability and accuracy of the mined association rule. There are two important measures for association rules, support and Confidence[5].

The support s of the rule is the prior probability of X and Y,

$$s = \sup(x \cap Y) = \frac{|X \cup Y|}{n}, and$$

The confidence c of the rule is the conditional probability of Y given X,

$$c = \frac{\sup(X \cup Y)}{\sup(X)} = \frac{|X \cup Y|}{|X|}$$

2. Weka :

Waikato Environment for Knowledge Analysis, called shortly WEKA, is a set of state-of-the art data mining algorithms and tools to in-depth analyses. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License[7].

3. Feature Selection :

Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy. In general; feature selection techniques can be split into two categories

- Filter method
- Wrapper method

In this paper we have used filter method for feature selection

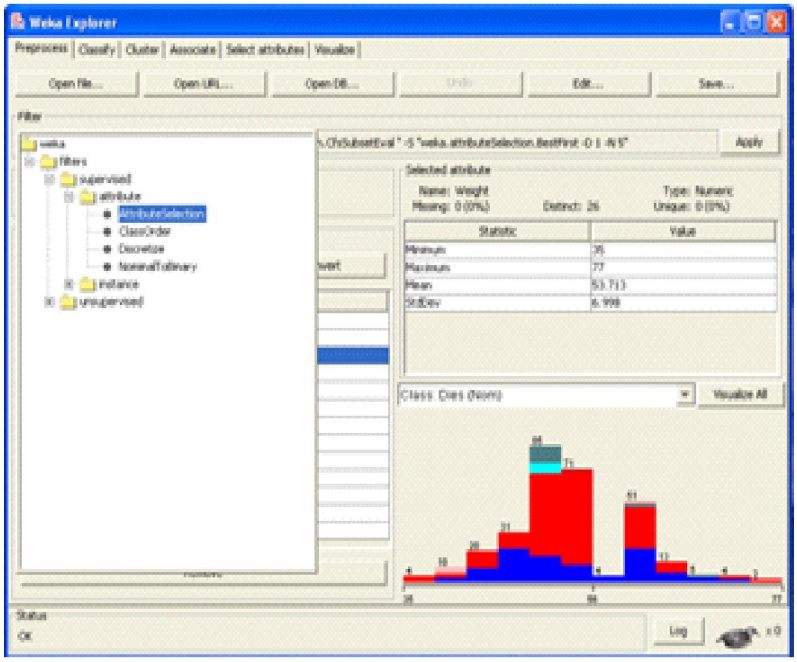

Fig. 1 WEKA explorer with Attribute Selection Method which is used in this research

4. Data Set Description :

Survey was carried out between Feb. 2008 to Jan 2009, and included interviews of 300 individuals. The survey was designed to gather information pertaining to the perceived health, diagnosed diseases, and general information such as name, address, age, sex, height, weight, education and income). The survey consisted of 20 questions. Some questions were to be answered yes or no, but generally respondents were provided with more options to answer the questions. The data was originally represented in excel data format in the form of two dimensional table consisting of 300 data points with each data point corresponding to the responses of an individual's, the dataset was converted into ARFF(Attribute Relation File Format) for effective and efficient usage WEKA system.

Data Reprocessing :

The process of data cleaning and preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. The researcher attempted to get ready the data according to the necessities of the selected data mining software such as WEKA and selected data mining algorithm. WEKA is multifunctional data mining software. The major data mining functions integrated in the software are data preprocessing, classification, association, clustering and visualizing input and output. Apriori is the only association rule algorithm implemented in Weka.

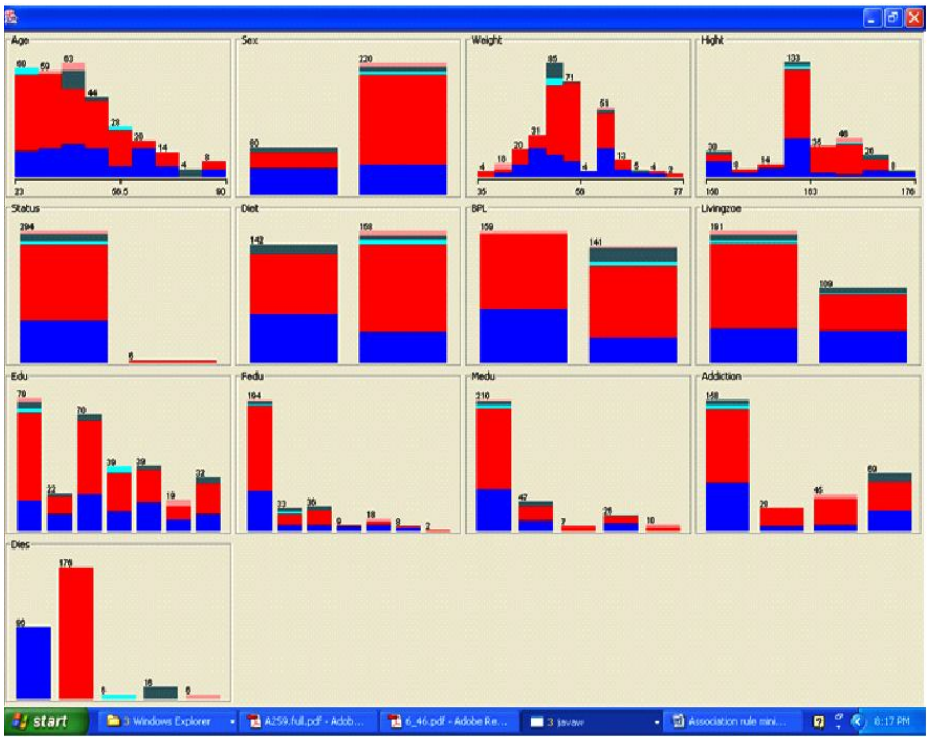

Fig. 2 Distribution of each Attributes of Dataset

5. Discovering Association Rules Using Weka:

WEKA generates association rules that have one or several output attributes. The strength of an association rule in WEKA is measured in terms of the rule's statistical significance, known as *support* and *confidence*. Support *s* is the percentage of transactions in *D* that contain $X \cup Y$, that is, the probability, $P(X \cup Y)$. Confidence *c* is the percentage of transactions in D containing X that also contain Y that is the conditional probability,[4] $P(X|Y)$. In this experiment we have used 80% confidence and 35% minimum support for good result[6].

5.1 Best Rules Found :

1. Fedu=unlit 194 ==> Medu=unlit 194 conf:(1)
2. Sex=m Fedu=unlit 134 ==> Medu = unlit 134 conf:(1)
3. BPL=y Fedu=unlit 131 ==> Medu = unlit 131 conf:(1)
4. Livingzoe=rural Fedu=unlit 128 ==> Medu = unlit 128 conf:(1)
5. Fedu=unlit Dies=ND 125 ==> Medu=unlit 125 conf:(1)
6. Diet=non_vegitarian Fedu=unlit 124 ==> Medu=unlit 124 conf:(1)
7. Fedu=unlit Addiction = Tobacco 116 ==> Medu=unlit 116 conf:(1)
8. Medu=unlit Dies=ND 129 ==> Fedu=unlit

125 conf:(0.97)
9. Medu=unlit Addiction=Tobacco 120 ==> Fedu=unlit 116 conf:(0.97)
10. Livingzoe=rural Medu=unlit 136 ==> Fedu =unlit 128 conf:(0.94)
11. Sex=m Medu = unlit 145 ==> Fedu=unlit 134 conf:(0.92)
12. Medu=unlit 210 ==> Fedu=unlit 194 conf: (0.92)
13. BPL=y Medu=unlit 143 ==> Fedu=unlit 131 conf:(0.92)
14. Diet=non_vegitarian Medu=unlit 136 ==> Fedu=unlit 124 conf:(0.91)
15. BPL=y 159 ==> Medu=unlit 143 conf:(0.9)
16. Diet=non_vegitarian 158 ==> Medu=unlit 136 conf:(0.86)
17. Livingzoe=rural Dies=ND 124 ==> Sex=m 105 conf:(0.85)
18. Dies=ND 176 ==> Sex=m 149 conf: (0.85)
19. BPL=y 159 ==> Fedu = unlit Medu=unlit 131 conf:(0.82)
20. BPL=y 159 ==> Fedu=unlit 131

6. Conclusion

In this paper we have applied data mining approach to discover association rules for population data. We have found 20 best rules. Rules can be interpreted. The interpretation of Rule 1: If father education uneducated then mother is also uneducated in 100% cases. This paper present the use of association rule mining in social scenario.

References

1. Agrawal A. T., Imielinsky, and Swami A., Mining Association rules between sets of items in large databases, In Proc. 1993 ACM – SIGMOD Int. Conf. Management of data (SIGMOD' 93), 207-216 (1993).
2. Agrawal R. and Srikant R., Fast Algorithms for Mining Association rules in large databases, Proceedings of the 20th International Conference on Very Large Data Bases, 487-499 (1994).
3. Chen M., Han J., Yu P.S., Data Mining: An Overview from Database Perspective, IEEE Transactions on Knowledge and Data Engineering, (8), (6), 310-315 (1996).
4. Han J. and Kamber M., Data Mining: Concept and Techniques. Morgan Kaufmann (2006).
5. Liu, B. Hsu, W., Ma, Y., Mining Association Rules with Multiple Minimum Supports," Proc. Knowledge Discovery and Data Mining Conf., 337-341 (199).
6. Omiecinski, E., Alternative Interest Measures for Mining Associations in Databases, IEEE Transactions on Knowledge and Data Engineering,(15), (1), 57-69 (2003).
7. Weka website: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/