The application of fuzzy association rules in the employment data mining of a higher vocational college

Laiquan Liu[1], Li Lei[2], Yanrui Lei[1*]

[1] *Hainan College of Software Technology, Fuhai Road No.128, Qionghai, Hainan, China*

[2] *Chongqing University of Arts and Sciences, Huachuang Road No.598, Yongchuan, Chongqing, China*

Received 1 March 2014, www.tsi.lv

Abstract

Data mining is able to extract potentially useful information from plentiful seemingly unrelated data. A high efficiency is therefore obtained using these useful data in work or study. Association rules mining is a significant branch in data mining. It mirrors the implicit relations among transactions in mass data. In addition, association rules can intuitively reflect the associations among item sets in data, and the relations are established according to the frequencies of the item sets appearing in data. This method, which explains its rules clearly and is easily to understand, therefore is different from the traditional statistical method. This research introduced and applied the mining algorithms of fuzzy association rules to the employment data analysis of a higher vocational college, in order to find significant association rules from numerous data and provide guidance for the education and employment in the future, therefore improving the employment rate further.

Keywords: Association rules, Data mining, Research, Application

1 Introduction

Data mining is an effective method to solve the problem of data rich but information poor currently. By using this method, potentially useful information can be discovered in mass data. Moreover, the relevant predication and discovering etc. of neglected information can be carried out using the discovered information. Data mining therefore presents broad application prospects.

Association rules mining is a significant branch in data mining. It is to discover the potential associated information in mass data. It was first reported in the data mining process of customers' transaction records in shopping malls [1]. There are no causal relationships in the results of association rules mining, and the results cannot be described using these relationships. Association rules mining mirrors the implicit relations among transactions in mass data. In addition, association rules can intuitively reflect the associations among item sets in data, and the relations are established according to the frequencies of the item sets appearing in data. This method, which explains its rules clearly and is easily to understand, therefore is different from the traditional statistical method. Association rules can intuitively express the relations among item sets (different values of variables) in data. The relations are not based on certain distributions and obtained using repeatedly iteration fittings of data in certain models. However, they are established according to the probabilities of the item sets appearing in data.

2 The definition of association rules

Let D={t_1,t_2,…,t_n} be a transactional database, T be any transaction set with an unique mark in D, and I={i_1,i_2,…,i_m} be a set composed of different items of number m. Each transaction ti (i=1,2,…,n) corresponds to an unique subset in I [2].

Definition 1. Item: Any element i in the set of I={i_1,i_2,…,i_m} is defined as a item.

Definition 2. Item set: In association analysis, a set containing none or multi-item is an item set. If an item set contains items of number k, it is called a k-item set. For example, {notebook computer, printer} is a 2-itemset. An empty set is an item set does not contain any items. If an item set X is a subset of a transaction T, that means the transaction T includes the item set X, which is denoted as $\mathbf{X} \subseteq \mathbf{T}$.

Association rules are implication expressions in the form of $\mathbf{X} \Rightarrow \mathbf{Y}$, where $\mathbf{X} \subseteq \mathbf{T}$, $Y \subseteq$ T , and $X \cap Y = \varphi$. The antecedent and the consequent of association rules are X and Y, respectively. Association rules mining is designed to find the implications which meet the set minimum support and confidence in mass data.

The support is applied to express the percentage of a rule in all the transactions in a database, and is a criterion to measure the importance of an association rule as well. The larger the support, the more important the association rule is in the whole database. The confidence is a measurement to determine the accuracy of an association

[*] ***Corresponding author*** e-mail: leiyanrui@139.com

rule. Generally, the association rules meeting the minimum support and confidence in the meantime are considered. If a rule is high support and low confidence, the rule presents low reliability; on the contrary, if a rule is low support and high confidence, the rule is seldom used.

Therefore, two threshold values, namely the minimum support and the minimum confidence, are set to guarantee the discovered association rules can be used in practice. These two values are employed to abandon the redundant and invalid rules.

Definition 3. Count. The number of the transactions containing item set X in the database D is called the count.

Definition 4. Support.: It is the ratio of the count of the item set X divided by the number of all the transactions in the database D. The support of item set X is denoted as sup(X), and the support of X $\Rightarrow$ Y is denoted as sup(X $\Rightarrow$ Y):

$$\sup(X \Rightarrow Y) = P(X \cap Y) \tag{1}$$

According to Definitions 3 and 4, in the case that the number of the transactions contained in the transaction database D is marked by |D|, the relation of the count and the support of item set X is expressed in the following formula:

$$\text{count} = \sup \times \text{D} \tag{2}$$

Definition 5. Confidence. The confidence of X $\Rightarrow$ Y is the specific value of the number of the transactions containing item sets X and Y at the same time and the number of the transactions merely containing item set X in the database D. The confidence of X $\Rightarrow$ Y is denoted as conf(X $\Rightarrow$ Y):

$$conf(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)} \times 100\% \quad . \tag{3}$$

Definition 6. Minimum support. According to the requirements, a threshold value is set which represents the minimum importance of the acceptable association rules, is denoted as minsup.

Definition 7. Minimum confidence. According to requirements, another threshold value is set. It displays the minimum confidence of the acceptable association rules and is denoted as minconf.

Definition 8. Strong association rule. The minsup and the minconf are set. If $\sup(X \Rightarrow Y) \geq \min\sup$ and $conf(X \Rightarrow Y) \geq \min conf$, the X $\Rightarrow$ Y is a strong association rule; otherwise, it is a weak one.

Association rules mining is aimed to discover all the strong association rules in D, the item sets corresponding to which must be frequent item sets. The process of association rules mining therefore begins with the discovering frequent item sets, then strong association rules are generated, and finally the rules are explained.

3 The process of association rules mining

Association rules mining is generally divided into and carried out as two subproblems [3].

3.1 DISCOVERING FREQUENT ITEMSETS

According to the minsup set by the decision maker, all the frequent item sets in the database D are found out. The frequent item sets refer to the item sets satisfying support as well as minsup. Since possibly there are inclusive relations among these frequent item sets, users need to find out the set of those frequent large item sets, which cannot be included in other frequent item sets. This is the basis of finally generating association rules as well.

3.2 DISCOVERING RULES

According to the set minconf, the rules with confidences not less than minconf were discovered in every maximum frequent item sets. The model of association rules mining is shown in Fig. 1.

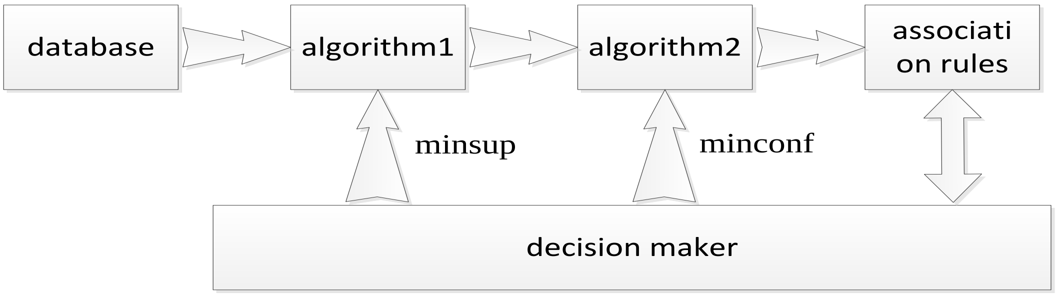


FIGURE 1 The model of association rules mining

4 Introduction of fuzzy association rules

The association rules are used most widely in market basket analysis. In a database, except items, there is numerical information relating to these items as well, such as the quantity and the price etc. of commodities. When first proposed, the association rules just considered the membership information, while ignored the numerical one. Researchers therefore focus on whether the numerical information is useful in further mining or not.

Generally, in reality, when carry out association rules mining of data which are not Boolean or categorical, researchers transform them to Boolean one. However, the transformation damages the edge data seriously, therefore influencing the mining results. Afterwards, the fuzzy concept was introduced in the association rules, and the problem of edge data was readily solved. The fuzzy association rules request the fuzzy concept is fuzzy as well as the membership functions. In the transformation, using different membership functions can result in different results, which influencing the mining results significantly [4].

Let I={I_1,I_2,…,I_m} be the attribute set of the database D. For any attribute I_i ($1 \le i \le m$), it can be divided into fuzzy attributes of number q_j using fuzzy membership functions. After the original numeric attributes being divided into fuzzy ones, the database D was transformed to fuzzy database D_f, the attribute set of which is I_f={ $I_1^1, I_1^2, \dots I_1^{q1}, I_2^1, I_2^2, \dots I_2^{q2}, \dots, I_m^1, I_m^2, \dots I_m^{qm}$,}, and the ranges of all the new attributes expand to [0,1].

Definition 9. The support of record i in fuzzy item sets X={x_1,x_2,…,x_p} $\subset I_f$ to the fuzzy item set X is defined by the following for mula:

$$SupT_i(X)=x_{1i}\wedge\ldots\wedge xpi \text{ or } SupT_i(X)=x_{1i}\times\ldots\times x_{pi}, \quad (4)$$

where x_{ji} represents the value of fuzzy item xj on the record i, and $x_{ji}\in[0,1]$ (i=1,2,…,n j=1,2,…,p).

Definition 10. The support of the whole fuzzy item set X={x_1,x_2,…,x_p} $\subset I_f$ to X is defined by the following formula:

$$\sup(X)=\frac{\sum_{i=1}^{n}\sup T_i(X)}{|D_f|}, \quad (5)$$

where $|D_f|$ represents the number of the transactions in the database. If the support of a fuzzy item set is not less than the set fuzzy minsup, X is a fuzzy frequent item set.

Definition 11. Similar to Boolean association rules, in the implication $X \Rightarrow Y$ in fuzzy association rules, X and Y indicate the antecedent and the consequent of the fuzzy association rules, respectively. Similarly, $X \subseteq I_f$, $Y \subseteq I_f$, and there are no relevant items from the same attributes in, $X \ne \varphi$, $Y \ne \varphi$, $X \cap Y \ne \phi$, and $I = X \cup Y$.

Definition 12. The $\sup(X \Rightarrow Y)$ and $conf(X \Rightarrow Y)$ of implication $X \Rightarrow Y$ in the fuzzy association rules are defined by the following formulas:

$$\sup(X \Rightarrow Y)=\sup(X \cup Y), \quad (6)$$

$$conf(X \Rightarrow Y)=\frac{\sup(X \cup Y)}{\sup(X)}, \quad (7)$$

where $X \subseteq I_f$, $Y \subseteq I_f$, and there are no relevant items from the same attributes in $X \ne \varphi$, $Y \ne \varphi$, $X \cap Y \ne \varphi$, and $I = X \cup Y$.

Similar to Boolean association rules, the minsup and the minconf are set by the decision maker prior to discovering fuzzy association rules. The fuzzy association rules mining is carried out in the following process: determining membership functions, establishing transactional database, discovering fuzzy association rules, and finally explaining discovered rules.

5 The application of fuzzy association rules in the employment data analysis of a higher vocational college

This research was on the basis of the accumulated relevant data of graduates of the vocational college in Hainan, China and analysed the employment trend and rules using data mining. The results are able to provide suggestions on vocational counsel and educational reform for the management and decision-making sections of the college, thus promoting the sustainable development of the college [5].

In this research, it is undoubtedly of great practical significance to apply data mining to the analysis of employment and educational reform, to discover the internal and hidden information from plenty of historical data using fuzzy association rules, and to employ the information to the decision-making of the college.

5.1 DATA PREPARATION

The collected employment information of 379 graduates from different majors of the vocational college in Hainan from 2012 to 2013 was employed in the research, and the relevant information such as name, gender, and birthday etc. was omitted. Then the following attributes were carried out association rules mining, including the average scores of common required courses, professional basic courses, and professional courses, majors, industries, income, and business natures etc.. Since the data size is large, this research merely shows partial data below, as shown in Table 1.

TABLE 1 The employment data of the vocational college in Hainan

Serial number	Average score of common required courses (ASCRC)	Average score of professional basic courses (ASPBC)	Average score of professional courses (ASPC)	Major	Industry	Income/ month	Business nature
001	86	82	84	software technology	IT	3800	private operated company
002	71	67	71	software technology	marketing	2700	private operated company
003	81	72	87	software technology	IT	4200	foreign company
004	81	87	68	software technology	education	2700	state-owned enterprise
…	…	…	…	…	…	…	…
376	65	72	69	electronic commerce	finance	2200	private operated company
377	71	67	71	electronic commerce	marketing	2100	private operated company
378	83	72	87	electronic commerce	education	3300	state-owned enterprise
379	81	83	89	electronic commerce	marketing	4100	foreign company

5.2 DISCOVERING ASSOCIATION RULES

1) The minsup and minconf are set to be 0.3 and 0.5, respectively. Then the clustering centres of the data were calculated using C means clustering algorithm. The centres are displayed in Table 2 [6]. The fuzzy database is not demonstrated due to its large size.

TABLE 2 Clustering centre of different attributes

Attribute	Clustering centre		
	1	**2**	**3**
ASCRC	58	71	81
ASPBC	51	73	83
ASPC	57	76	83
Income	2100	3300	3900

2) The Counts of all the fuzzy items were calculated and the fuzzy items with Counts not less than 35 were classified into frequent 1-itemsets L_1. Then C_2 was generated by connecting the item sets L_1, and the fuzzy items corresponding to the same attributes were not connected. Afterwards L_2 was generated by C_2 and C_3 was generated by connecting L_2. In the case that $C_4 = \phi$, the mining is therefore finished. The generated association rules are indicated in Table 3 [7].

The association rules in Table 3 indicate that most of the graduates with medium scores of professional courses and high scores of professional basic ones work in the industries not related to their majors; most of the graduates with high scores of professional and professional basic courses work in the industries related to their majors; and the incomes of the graduates with high scores of common required, professional and professional basic courses are generally at a high level.

The above data indicate that, in order to work in the industries related to their majors after graduation, students require to concentrate their attentions on all the following courses, including common required, professional basic, and professional courses, so that they can develop in an all-around way; meanwhile, if some students are interested in other majors, proper suggestions need to be given to improve their relevant professional quality, so that they are ready for the employment in the future. The information provides guidance and references for the training scheme of the higher vocational college. And it is useful in the training of the applied talents needed for society and improving the employment rate of students.

TABLE 3 The discovered association rules

Association Rules	Support (%)	Confidence (%)
{ASCRC.high, ASPC.medium} $\Rightarrow$ not related	33.4	72.6
{ASCRC.high, ASPC.high} $\Rightarrow$ related	34.7	64.1
{ASCRC.high, ASBPC.high, ASPC.high} $\Rightarrow$ high income	37.5	73.6

6 Conclusions

The research employed the fuzzy association rules algorithms in the analysis of the employment of the higher vocational college and discovered the relationships among the scores and the employment attributes. According to this method, the fuzzy C means clustering algorithm was used to cluster the quantitative attributes, and then the clustering centres were mined using the association rules mining. However, the research needs to be perfected in some aspects as well. For example, the research discovered all the strong association rules, of which some with high support and confidence are not valuable in practical applications. In addition, whether the discovered quantitative association rules are valuable in application or not needs to be verified as well.

References

[1] Agrawal R, Imielinski T, Swami A 1993 Mining association rules between sets of items in large databases, in *Proc. ACM SIGMOD Int. Conf. Management of Data*Washington DC **5** 207-16

[2] Han J W 2008 *Data Mining Concepts and Techniques* Beijing: China Machine Press 146-76 *(in Chinese)*

[3] Au W H and Chart K C C 2003 *IEEE T. Fuzzy Syst.***11**(2) 238-48

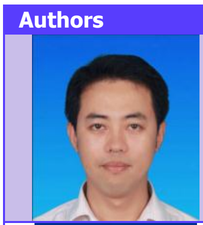
[4] Yan P and Chen G Q 2004 *Fuzzy Syst. Math.* **18**(1) 279-83

[5] Ru J L 2013 *The research and implementation of algorithm in the employment of higher vocational colleges in the mining of association rules* Chengdu: Univ. Electr. Sci. Technol. China 2-6 *(in Chinese)*

[6] Wen H 2013 Knowledge map mining of financial data *J. Tsinghua Univ. (Science and Technology)* **1**(1) 68-76 *(in Chinese)*

[7] Zhang X P 2013 Application of quantitative association rules in college employment information data *Comput. Technol. Dev.* **23**(11) 199-203*(in Chinese)*

Authors

	**Laiquan Liu, born in August 14, 1979, Shaanxi Province of China** **Current position, grades:** Hainan College of Software Technology, associate professor, senior engineer **University studies:** Shanxi Normal University (Bachelor of Science in Educational Technology), 2003; Tianjin University of Science & Technology (Master's degree in Control engineering), 2012 **Scientific interest:** vocational education, multimedia application and data mining **Publications:** more than 15 **Experience:** Hainan Academy on Computers of China (2008-)
	Li Lei, born in October 16, 1979, Shaanxi Province of China **Current position, grades:** Chongqing University of Arts and Sciences, lecturer **University studies:** Shanxi Normal University (Bachelor of Science in Computer science and technology), 2003; Chongqing University (Master's degree in Computer technology), 2013 **Scientific interest:** web application development and data mining **Publications:** more than 5
	Yanrui Lei, born in December 12, 1980, Shaanxi Province of China **Current position, grades:** Hainan College of Software Technology, lecturer, engineer **University studies:** Shanxi Normal University (Bachelor of Science in Computer science and technology), 2003; Sun Yat-Sen University (Master's degree in Software engineering), 2013 **Scientific interest**: web application development and data mining **Publications:** more than 15 **Experience:** Hainan Academy on Computers of China (2008-)