A RELIABILITY EVALUATION SYSTEM OF ASSOCIATION RULES

Jiangping Chen[a] , Wanshu Feng[a], Minghai Luo[b,*]

[a]School of Remote Sensing and Information Engineering ,Wuhan University, Wuhan,Hubei,430079,China
[b]Wuhan Geomatics Institute,Wuhan,Hubei,430022, China

Commission II, WG II/3

KEY WORDS: Association rules, Evaluation, Reliability, Accuracy, Completeness, Consistency

ABSTRACT:
In mining association rules, the evaluation of the rules is a highly important work because it directly affects the usability and applicability of the output results of mining. In this paper, the concept of reliability was imported into the association rule evaluation. The reliability of association rules was defined as the accordance degree that reflects the rules of the mining data set. Such degree contains three levels of measurement, namely, accuracy, completeness, and consistency of rules. To show its effectiveness, the "accuracy-completeness-consistency" reliability evaluation system was applied to two extremely different data sets, namely, a basket simulation data set and a multi-source lightning data fusion. Results show that the reliability evaluation system works well in both simulation data set and the actual problem. The three-dimensional reliability evaluation can effectively detect the useless rules to be screened out and add the missing rules thereby improving the reliability of mining results. Furthermore, the proposed reliability evaluation system is applicable to many research fields; using the system in the analysis can facilitate obtainment of more accurate, complete, and consistent association rules.

1. INTRODUCTION

Association rule mining is an important research topic in data mining because it aims to define the effective, reliable, understandable, and interesting rules from a large database, which could assist people in analyzing and decision making (Han and Kamber, 2006; Pauray and Chen, 2004). Thus, the evaluation of association rules is highly significant because it directly affects the reliability of the output rules obtained by mining.

Currently, studies on association rules mainly focus on the efficiency improvement of the mining algorithm; discussion regarding the evaluation methods of mining result is relatively minimal (Song, Zhai and Gao, 2007). The existing evaluation method of association rules can be divided into two types, namely, objective evaluation and subjective evaluation (Ji and Sun, 2004). Objective evaluation depends on the specific structure of association rules and the mining data; it judges the association rule using quantitative values calculated by statistical methods. Common objective evaluation contains many statistical indicators, such as support, confidence (Hannu, 1996), and lift (Berry and Linoff, 1997). Subjective evaluation assesses the rule using subjective factors instead of the data in the database. These factors mainly reflect the user's participation and the knowledge fusion of the field, such as the potential usefulness, conciseness, and so on (Lou, Jiang, and Tian, 2003). However, objective evaluation of association rules only relies on the structure of the data and does not consider the connection among the rules and subjective evaluation requires a mass of prior information offered by users and experts, which relies on a controversial assessment of indicators (Zhu and Hu, 2007). These commonly used association rule evaluation systems mainly work by excluding useless rules via some threshold limits, rather than emphasizing the reliability measurement of the association rules.

In this paper, the concept of reliability was imported into the association rule evaluation. The concept of reliability stems from the description of the product's reliability, which refers to the capability that enables the product to complete the required function under a\specified condition and within the specified time (Sinha, 1986). Current research on reliability and association rule tends to consider rule mining based on statistical data, such as the product's basic reliability (Jayakrushna and Ashok, 2015) and useful life (Xu, 2008); or assess the data quality for mining using some reliability methods (Katja and Marc-Thorsten, 2011; Sandro, Benny and Ian, 2005). However, current research rarely provides objective and quantitative judgment on the reliability of the association rule. In this paper, the reliability of association rule was defined as the reliability acquired by the association rule under a specified condition that reflects the correlation among the data in the database. The definition includes three levels of measurement: 1) accuracy of association rule to describe data (Accuracy); 2) integrity of association rule to cover the correlation among the data (Completeness); and 3) consistency of the distribution of association rule in the data (Consistency) (Shi et al., 2012). The definition, measurement, and significance of the three levels of reliability evaluation were proposed in the text. The "accuracy-completeness-consistency" three-dimensional reliability evaluation system can serve as a reference and tool, for evaluating the reliability of association rules in various fields.

2. RELIABILITY EVALUATION METHODS OF ASSOCIATION RULES

2.1 Reliability evaluation indicator

2.1.1 Accuracy: Accuracy of association rules refers to the degree of accuracy of the association rules that represent the correlation among attributes in the transaction database. Accuracy integrates the assessment of the following three levels: 1) usefulness of the association rules representing the correlation among the data (support); 2) certainty of association rules reflecting the correlation (confidence); 3) correctness of the association among the structures of association rules (lift). Thus, the accuracy of rules was defined as a function of three objective indicators, namely, support, confidence, and lift. Considering that these three indicators were calculated based on the specific structures of the rules of statistical methods (Mennis

and Guo, 2009), they are capable of avoiding the influence of artificial opinion and can represent the degree of accuracy of the association rules.

Given that the dimensions for support, confidence, and lift are inconsistent, we defined the accuracy function of association rule as a geometric weighted mean of the three indicators as follows:

$$\text{Accuracy} = \text{Support}^{k_1} \times \text{Confidence}^{k_2} \times \text{Lift}^{k_3} \ , \quad (1)$$

where $k_i \geq 0.(\mathrm{i} = 1,2,3)$

$\sum_{i=1}^{3} k_i = 1$.

k_1, k_2 , and k_3 denote the weights of support, confidence, and lift in the accuracy function, respectively. By default, $k_1 = k_2 = k_3 = 1/3$ exists, which indicates the equal importance of the three indicators (Su, You, and Yang, 2004). Depending on the characteristics of the research fields and mining data, the weight of the three indicators can be adjusted accordingly.

2.1.2 Completeness: Completeness of the association rules is a measurement of the integrity degree calculated by the association rules from the research data set, which expresses the correlation in the database. The key to evaluate completeness is to judge whether any missing rules exist in the association rule mining results, that is, whether the newly presented rules acquired by comparing the mining results from different data sets with the study rule set are the missing correct rules or not. In this paper, we added the missing rules to the study rule set using the concept of novelty. Figure 1 shows the process of completeness evaluation.

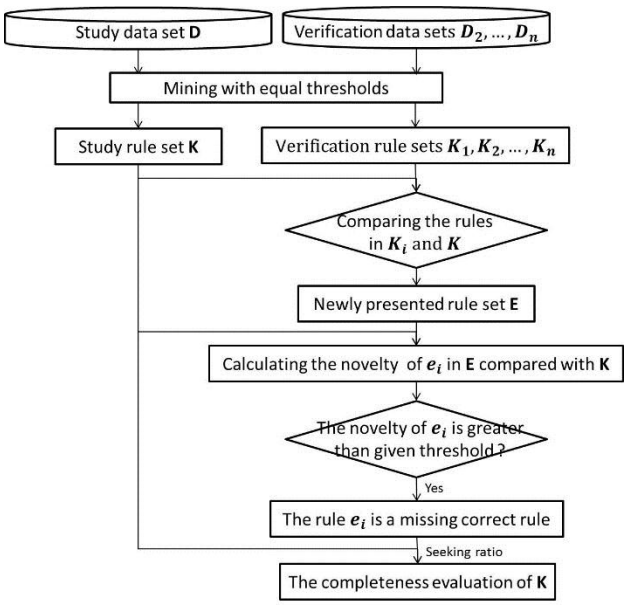

Figure 1. The process of completeness evaluation

The novelty of a newly presented rule reflects the level of difference between the new rule and the study rule set (Qi, 2004). If the value of novelty for a new rule was high, the rule was considered for omission in the initial mining and added into the study rule set.

The mining results of the study data set are regarded as rule set **K**; the number of rules in **K** is |**K**|. By assessing the novelty of rules in the newly presented rule set **N**, the rules with higher novelty constitute the missing rule set **M**; the number of rules in **M** is |**M**|. The completeness of the association rules is the ratio of the number of association rules in the origin study rule set and completed rule set.

$$\text{Completeness} = \frac{|\boldsymbol{K}|}{|\mathbf{K}| + |\boldsymbol{M}|} \ . \quad (2)$$

2.1.3 Consistency: Consistency of association rules is derived through the assessment of some common rules appearing in the mining results of different data sets and the uniformity of support distribution for these rules in different data sets. If a rule appeared in the mining results of the study data set as well as in all the appended different data sets, and if the frequency distribution of the rule in these data sets is consistent, the consistency of the rule can be concluded as good. If the consistency of the rule set is poor, meaning that the distribution of study rules in the data sets is uneven, then the data sets should be appropriately split to obtain more reliable and complete association rules. The result of consistency evaluation includes two parts, namely, the Result Consistency (RC) and the Data Consistency (DC).

RC is the proportion of rules appearing in the mining results of different data sets, which constitute the rule set **C**; and the number of rules in **C** is |**C**|. The consistency of results is the ratio of the number of association rules in the common rule set **C** and study rule set **K**.

$$\text{RC} = \frac{|\boldsymbol{C}|}{|\boldsymbol{K}|} \ . \quad (3)$$

DC is the standard deviation value of the support distribution for these common rules in different data sets. The support for rule C_i in the study data set is S_i and S_{i1}、S_{i2}、…S_{in} are assigned to the other different data sets. The DC can be calculated using the following steps:

(1) Computing v_i :

$$v_i = |S_{i1} - S_i| + |S_{i2} - S_i| + \ldots + |S_{in} - S_i| \ . \quad (4)$$

(2) Standardizing v_i ; the maximum of v_i for every single rule in C is v_{max} , and the minimum is v_{min} :

$$v_i^{'} = \frac{v_i - v_{min}}{v_{max} - v_{min}} \ . \quad (5)$$

(3) Judging the DC of association rules:

$$\text{DC} = 1 - v_i^{'} \ . \quad (6)$$

According to the definition of RC and DC, the consistency of rules is a combination of the two results:

$$\text{Consistency} = k_1 \times \text{RC} + k_2 \times \text{DC} \ , \quad (7)$$

where $k_1, k_1 \geq 0$

$k_1 + k_2 = 1$.

k_1 and k_2 are the weights of indicators RC and DC in the

consistency evaluation. By default, $k_1 = k_2 = 1/2$ exists. According to the different demands in actual problems, the value of weights should be adjusted.

2.2 "Accuracy-completeness-consistency" reliability evaluation system

We established the framework of the reliability evaluation system according to the definitions and formulas of accuracy, completeness, and consistency in association rules (see Figure 2).

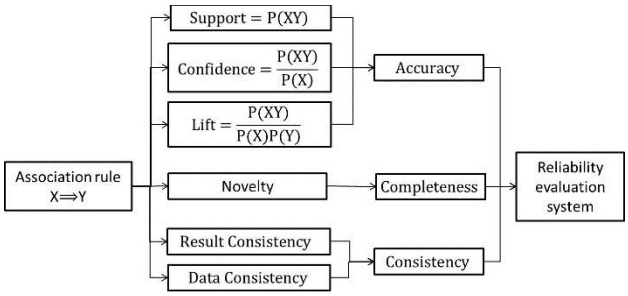

Figure 2. Reliability evaluation system

The accuracy evaluation can realize the assessment of the frequency, intensity, and correlation of association rules. The evaluation is capable of screening out the rules that did not meet the requirements. Furthermore, it ensures that the result rule sets are the most accurate expression of the correlation among the mining data. According to the mining result, some missing rules can be acquired via the completeness evaluation. Furthermore, the degree of association rules covering the correlation in the data set can be measured. According to the consistency evaluation, the continuity of rules appearing in the data mining can be evaluated, which is important in the decision of the application strength of association rules. The "accuracy-completeness-consistency" reliability evaluation system realizes the multi-level assessment of rules and is extremely necessary in improving the reliability of the association rules.

3. SIMULATION EXPERIMENT OF RELIABILITY EVALUATION

3.1 Reliability evaluation of association rules based on simulated data

We used a simulated market basket data set to verify the reliability evaluation methods of association rules proposed in Chapter 2. Table 1 shows a recording statistics of 500 customers of a supermarket, where 1 represents a customer purchasing goods and 0 represents none. By mining and evaluating the customer consumption data set, the reliability of the basket rules for the supermarket can be assessed to evaluate the shopping pattern in the supermarket.

	Apple	Avocado	Baguette	Wine	…
1	1	1	0	1	…
2	1	0	1	1	…
…	…	…	…	…	…
500	1	1	1	1	…

Table 1. Recording statistics of customer consumption in a supermarket

3.1.1 Accuracy of association rules: To measure the accuracy of the rules in relation to the supermarket, the classic *A*priori algorithm (Agrawal, Imielinski, and Swami, 1993) was adopted to mine the basket data set. To derive a set of rules that meets the threshold requirements, the threshold setting was established based on the values of minimum support, confidence, and lift that are greater than 50%, 80%, and 1, respectively. Table 2 shows the accuracy results for the study rule set **K** obtained by calculating the accuracy of each rule with the corresponding values of support, confidence, and lift.

	Antecedent	Consequence	Accuracy
1	Coke_1	Wine_1	0.85
2	Ice cream_1	Coke_1	0.88
3	Steak_1	Baguette_1	0.88
4	Apple_1	Steak_1	0.82
5	Ice cream_1	Sardine_1	0.79
6	Beer_0	Steak_1	0.79

Table 2. Study association rule set **K** and its accuracy evaluation

In descending order of support, the rule of the highest frequency in the data is "Coke_1→ Wine_1." The result shows that 66% of customers who simultaneously bought both Coke and wine, and 93% of the customers who bought Coke would buy wine; thus, the purchase of Coke has a positive correlation with the purchase of wine. Consequently, the accuracy of the rule "Coke_1 → Wine_1 " is 0.85. The accuracy of the study association rule set is 0.84 according to the mean.

3.1.2 Completeness of the association rules: In the same supermarket basket data, a new data set that also contains 500 records was appended. The appended data set was mined with *A*priori algorithm under the same threshold value (the minimum limits for support, confidence, and lift are 50%, 80%, and 1, respectively). Table 3 lists the 12 rules of the newly presented market basket data that meet the threshold limit.

	Antecedent	Consequence	Sup	Conf	Lif
1	Ice cream_1	Wine_1	0.64	0.94	1.01
2	Artichoke_1	Wine_1	0.64	0.94	1.01
3	Ice cream_1	Coke_1	0.63	0.89	1.27
4	Steak_1	Baguette_1	0.59	0.87	1.27
5	Ham_1	Turkey_1	0.58	0.80	1.10
6	Apple_1	Steak_1	0.56	0.82	1.06
7	Corned beef_1	Steak_1	0.55	0.85	1.10
8	Wine_1,Turkey_1	Ham_1	0.54	0.80	1.11
9	Avocado_1	Artichoke_1	0.53	0.81	1.20
10	Beer_0	Turkey_1	0.52	0.85	1.17
11	Beer_0	Steak_1	0.50	0.85	1.09
12	Beer_0	Apple_1	0.50	0.83	1.19

Table 3. Mining result of the newly presented market basket data

Compared with the study rule set listed in Table 2, the newly presented rules in the appended rules were obtained as rule set **N**. Calculating the novelty of every single rule in **N** is relative to the study rule set **K**. The novelty threshold was set as 3 to remove the rules with lower novelty and retain the rules with

higher novelty as the interesting rules omitted in the basket data mining, which constitute the missing rule set **M** (see Table 4).

	Antecedent	Consequence	Novelty
1	Ham_1	Turkey_1	4
2	Wine_1,Turkey_1	Ham_1	4
3	Avocado_1	Artichoke_1	4
4	Artichoke_1	Wine_1	3.67
5	Beer_0	Turkey_1	3.67
6	Beer_0	Apple_1	3.67
7	Corned beef_1	Steak_1	3.33

Table 4. Missing rule set **M** in market basket data mining

According to the rules listed in Table 2 and Table 4, the number of rules in study rule set |**K**| is 6, and the number of rules in missing rule set |**M**| is 7. Subsequently, the completeness of the basket association rules can be calculated as follows:

$$\text{Completeness} = \frac{|\boldsymbol{K}|}{|\mathbf{K}| + |\boldsymbol{M}|} = \frac{6}{13} = 0.46 .$$

3.1.3 Consistency of association rules: Consistency of basket rules includes the result consistency (RC) and the data consistency (DC).

(1) Result consistency (RC)

Comparing the study rule set **K** with the appended rule set **E**, some rules appearing in both basket rule sets can be defined as the common rule set **C**. Table 5 lists the statistics of support distribution of these common rules in both study data set and appended data set.

	Antecedent	Consequence	Support	
			K	**C**
1	Ice cream_1	Coke_1	0.62	0.63
2	Coke_1	Ice cream_1	0.62	0.59
3	Apple_1	Steak_1	0.58	0.56
4	Beer_0	Steak_1	0.51	0.50

Table 5. Support distribution of rules of **C** in different data sets

Table 5 shows that the number of rules in common rule set |**C**| is 4; thus, the RC of rules can be calculated as follows:

$$\text{RC} = \frac{|\boldsymbol{C}|}{|\boldsymbol{K}|} = \frac{4}{6} = 0.67 .$$

(2) Data consistency (DC)

According to the support distribution of common rules in different data sets, Table 6 shows the calculation of DC.

	Antecedent	Consequence	v_i	v_i'	DC
1	Ice cream_1	Coke_1	0.01	0	1
2	Coke_1	Ice Cream_1	0.03	1	0
3	Apple_1	Steak_1	0.02	0.5	0.5
4	Beer_0	Steak_1	0.01	0	1

Table 6. DC of the association rules

The DC is 0.63 according to the mean of values in Table 6, and the consistency of study rule set is 0.65.

3.2 Analysis of the reliability evaluation results

3.2.1 Analysis of accuracy: According to the accuracy evaluation, the accuracy of association rules obtained from basket data is 0.84. The value indicates that the obtained rule set has high frequency, intensity, and relevance in the database and has a good expression of the correlation among the data.

In the course of the accuracy evaluation of the basket data set, the values of the threshold limit for support, confidence, and lift are 50%, 80%, and 1, respectively. By screening out the association rules that failed to meet the threshold requirement, the accuracy of the mining results was effectively improved. Table 7 lists some excluded rules.

	Antecedent	Consequence	Sup	Conf	Lif
1	Avocado_0	Whiskey_1	0.31	0.81	1.33
2	Sardine_1	Ice Cream_1	0.55	0.77	1.12
3	Corned beef_1	Wine_1	0.52	0.81	0.98
4	Coke_1	Sardine_0	0.42	0.68	1.32
…	…	…	…	…	…

Table 7. Examples of excluded rules

To illustrate, the 3rd rule is used as example ("Corned beef_1⇒Wine_1)." Support and confidence of the rule indicate that 52% of the customers would buy corned beef and wine simultaneously, and 81% of the customers who bought corned beef would buy some wine as well, which seems to be a strong rule. However, according to the statistics of basket data set, the maximum probability of the event "Buy wine" is 84%, whereas the rule showed that only 81% of customers who bought corned beef would buy wine. Thus, the event "Buy corned beef" will lower the possibility of "Buy wine," which denotes that the two are actually negatively correlated. The indication of the rule, "Corned beef_1⇒Wine_1," which means that the customer who bought corned beef tends to buy wine simultaneously, is apparently incorrect and should be screened out of the study rule set. The exclusion of association rule based on the threshold limit for three accuracy indicators effectively guarantees the accuracy of the rule to meet the user's needs.

3.2.2 Analysis of completeness: According to the completeness evaluation, the completeness of association rules obtained from basket data is 0.46, indicating that the obtained rule set cannot fully cover the correlation in the basket data set.

In the course of the completeness evaluation, the association rules obtained under the same threshold from the study data set and a newly-presented shopping record were contrasted. Consequently, the newly presented association rules can be captured to calculate the novelty compared with the study rule set and consequently determine whether the new rules are missing rules in the basket data or not. Table 2 and Table 4 show the study rule set and missing rule set, respectively. Considering the insufficient basket data and various mining attributes (20 types of goods), the completeness of study rule set is not so good. Due to the diversity of goods, the newly presented rules tend to have a higher novelty value compared with the study rule set. For example, the items "Corned beef_1" and "Steak_1" belong to different language variables. Furthermore, the novelty value between them is at its highest, which denotes that a large number of new rules will be retained and result in low completeness.

In the actual supermarket record, the types of goods bought by customers are multifarious. The observation indicates that the number of attributes for the association rule mining is more than 20. "Buy tungsten filament bulbs" and "Buy LED bulbs," or "buy the dishes of different brands" would be classified as different language variables in mining. The dispersion of data increases the difficulty in finding strong rules on the level of detail and in acquiring a more complete shopping pattern for the supermarket.

To improve the completeness of the supermarket mining result, the idea of conceptual hierarchy can be used in the mining association rules (Han, Cai, and Cercone, 1992). The goods can be divided into different categories and levels according to types, and the mapping sequence can be established based on the underlying concept (concrete concept) and the high-level concept (abstract concept). By generalizing specific goods into a higher level, the association rules can be obtained on multiple levels. Furthermore, optimization of rule completeness is realized.

3.2.3 Analysis of consistency: According to the consistency evaluation, the consistency of association rules obtained from basket data is 0.65, whereas that of RC is 0.67, and the DC is 0.63.

If the common rule set appearing in the mining results of different shopping records has high proportion, and the support distribution of these common rules in different data sets is relatively consistent, the rule set considered as the shopping pattern for the supermarket is believed to have significant consistency. The consistency evaluation of the basket data shows that 67% of the association rules are completely consistent in the mining results of different data sets, and the data consistency of the distribution of these common rules in various data sets is 63%. Therefore, the consistency of the rule set for the supermarket is relatively good, and the shopping pattern shown in Table 5 should be fully considered in the supermarket management.

4. RELIABILITY EVALUATION OF THE LIGHTNING ASSOCIATION RULE IN HUBEI PROVINCE

4.1 Data introduction

Lightning activity is an important factor that significantly influences the social production and daily life. In recent years, lightning caused several accidents and property damage in relation to human and animal safety, construction, electricity, telecommunications, forestry, and other fields. Thus, exploring and evaluating the association rule regarding lightning to obtain a more accurate and complete lightning association rule set is highly important.

To further verify the reliability evaluation system described above, the lightning monitoring data of the main transmission line corridors in Huangshi, Wuhan, Xianning, and Yichang were extracted from the Hubei lightning monitoring system database. Then, the lightning data sets for the four cities were combined with meteorology, terrain, and vegetation factors. Furthermore, the application of reliability evaluation methods was explored in the lightning activity rule mining. The experimental data include two parts, namely, the lightning data and the multi-source data.

(1) Lightning data

Summer is period of the year that has the most frequent lightning activity. Summer is also the most important period for lightning accident preservation and mitigation work. Thus, the July and August 2014 lightning data of the four cities were selected as the study data set (see Table 8).

Month	Lightning monitoring record (unit: item)			
	Huangshi	Wuhan	Xianning	Yichang
7	2513	832	3044	3701
8	1470	1078	945	1928
7,8	3983	1910	3989	5629

Table 8. Lightning monitoring data of Huangshi, Wuhan, Xianning, Yichang

(2) Multi-source data

The influence of terrain, meteorology, and vegetation was considered. The multi-source fusion (see Table 9) for lightning data can be obtained for the association rule mining based on the terrain data of Hubei Province (elevation, slope, and aspect), the meteorological data (precipitation, temperature, barometric pressure, relative humidity, etc.) and the vegetation coverage data. Among them, the reference of vegetation is shown in Table 10, and the classifications of aspect and direction of maximum wind speed are determined according to the angle (see Figure 3).

Attribute	Unit	Attribute	Unit
Lightning		**Meteorology**	
current amplitude	kA	direction of maximum wind speed	coding
Vegetation		maximum wind speed	m/s
vegetation	coding	average wind speed	m/s
Terrain		precipitation	mm
elevation	m	average temperature	℃
slope	°	minimum temperature	℃
aspect	coding	average vapor pressure	hPa
		average barometric pressure	hPa
		average relative humidity	%

Table 9. Attributes of the lightning mining

wood land	farm land	building land	bare land	water area	grass land	un named
1	2	3	4	5	6	7

Table 10. Reference of vegetation

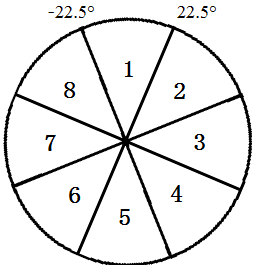

Figure 3. The classifications of aspect and direction of maximum wind speed

4.2 Reliability evaluation of the lightning rule

The four multi-source lightning data sets in July and August in Huangshi, Wuhan, Xianning, and Yichang were set as the study data sets. The data sets of July and August were separated to verify the association rule mining results for the four cities, and the reliability of the four study rule sets was explored.

The attributes of lightning data contain a large number of detailed information, if association rule mining is applied on these detailed data, many confusing rules may be gained. Before rule mining, the lightning data needs to be generalized, which aims at replacing the lower level of the object with a more abstract concept, so as to carry out the association rule mining at a higher level and obtain more broadly and meaningful association rules. The data generalization algorithm used in this study is FCM (Dunn, 1973), and it was applied in the 11 attributes of the lightning data set, except the attributes whose unit is coding (vegetation, aspect and the direction of maximum wind speed).

The Apriori algorithm was applied to acquire the association rules with the generalized lightning data for the four cities, whereas the minimum limits of support, confidence, and lift are 50%, 80%, and 1, respectively. Given that the topic of research is lightning activity, the rules whose consequence is "current amplitude" were extracted as the study rule sets. Two verification rule sets were used for the experiment, namely, a rule set for July and a rule set for August obtained under the same thresholds. With three rule sets of July and August combined, the separated rules for July and August of Huangshi, Wuhan, Xianning, and Yichang can be obtained.

In the course of the reliability evaluation of the lightning association rule, the measurement of the accuracy and consistency followed the default settings, whereas the threshold for novelty in completeness evaluation is 4/3. Using the lightning data of Wuhan as an example, Figure 4 shows the study rule set for July and August combined according to the Apriori algorithm. In the completeness evaluation for Wuhan, the study rule set was compared with the two rule sets containing the individual rules for July and August, respectively. The newly presented rules for July and August form the newly presented rule set. The novelty of the rules in the newly presented rule set was calculated. As a result, the rules with high value of novelty were retained as the missing rule set of summer lightning activity in Wuhan. The completeness can be obtained according to Equation 2. Furthermore, in the consistency evaluation for Wuhan, the study rule set should further undergo comparison with the two verification rule sets . Furthermore, the common rules in the three rule sets can be captured to calculate the RC and DC.

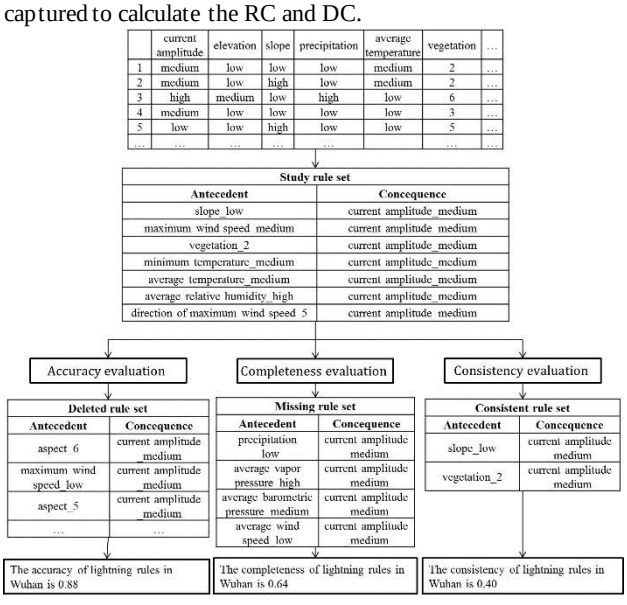

Figure 4. Reliability evaluation of the lightning association rule in Wuhan

4.3 Overall analysis of reliability evaluation results of summer lightning activities in Hubei Province

Table 11 shows the evaluation of the reliability of association rules of the summer lightning activities in Huangshi, Wuhan, Xianning, and Yichang, as well as the reliability indicators for the four cities.

City	Reliability evaluation system		
	Accuracy	Completeness	Consistency
Huangshi	0.86	0.80	0
Wuhan	0.88	0.64	0.40
Xianning	0.70	0.75	0.44
Yichang	0.87	0.62	0.38

Table 11. Reliability evaluation results of the summer lightning activities for the four cities

By analyzing the reliability evaluation results of the summer lightning activities in Huangshi, Wuhan, Xianning, and Yichang, the following conclusions were obtained.

(1) The accuracy results of association rule sets for four cities were at a high level. With the combination of multi-source data, which specifically indicate the meteorology, terrain, and vegetation data, the mining on lightning data can reflect effectively the correlation in the lightning activity in summer. Furthermore, the study rule set can provide support for the disaster prediction and prevention of summer lightning, and effectively can decrease the threat of lightning to production and life.

(2) The completeness results of the association rule sets for the four cities were poor. Based on the consistency of the lightning association rules in summer, the data in July and August were separated to enable the calculation of the completeness of the study rule set from the comparison of the combined July and August rules and two verification rule sets. Given that some differences exist between the meteorological characteristics of July and August, the corresponding association rules were relatively different, which lead to a poor outcome of the completeness evaluation. According to the meteorological characteristics of the lightning experiment, the mining of lightning rules in summer should be divided into two parts, namely, general and local parts. The different emphases of individual lightning activities in July and August should be considered in the association rule mining. The targeted summer lightning prevention and control measures can be developed based on the different characteristics of July and August.

(3) According to an overall analysis of reliability evaluation results for four cities, the comprehensive reliability of rules in Xianning was the best, given that the accuracy, completeness, and consistency for Xianning were all relatively good. Therefore, the relevant planning and deployment of lightning protection work in Xianning can be implemented according to the association rules. The completeness of Huangshi is the highest, but its consistency is 0, indicating that the language variables of antecedent (influence factors) are largely the same in the study rule set and in the two verification rule sets. However, the characteristics of July and August vary significantly.

CONCLUSIONS

This study proposed a reliability evaluation system for the objective and quantitative evaluation of the reliability of the association rules. Based on the principle of association rule mining and definition of reliability, the accuracy, completeness,

and consistency of association rule were described. To show its effectiveness, the reliability evaluation system was applied to two extremely different data sets, namely, a basket simulation data set and a multi-source lightning data fusion.

(1) By evaluating the association rules from the simulated basket data set using the reliability evaluation system, the useless rules can be effectively detected for screening out. Moreover, the missing rules can be added into the rule set, which can ensure the accuracy of association rules. Furthermore, if the mining data set has a large number of attributes that can be layered, the result of the completeness evaluation is generally ineffective, given that some similar attributes will be defined as high novelty. According to the characteristics of the basket data, the improvement measures, or the mining association rules based on hierarchy, were put forward. By measuring the consistency of the association rules, the shopping patterns obtained by association mining of the supermarket were applied.

(2) A practical data mining problem was studied. The summer lightning monitoring data sets of Huangshi, Wuhan, Xianning, and Yichang, which combined the terrain data of Hubei Province (elevation, slope, aspect), the meteorological data (precipitation, temperature, barometric pressure, relative humidity, etc.), and the main transmission line corridors vegetation data fusion, were adopted for the mining association rules of the lightning activities. By evaluating the reliability of association rules and analyzing the distribution characteristics of the reliability indicators in the four cities, the rules of the summer lightning activities were obtained, which were more accurate and complete than the original research rule set. Finally, suggestions for the exploration of the lightning activities, as well as for the forecast and prevention of lightning damage were made based on the analysis of reliability.

The results show that the reliability evaluation system works well in both simulation data and the actual problem. Through the accuracy evaluation methods, the inaccurate rules can be found and screened out. Such action improves the precision of association rules in describing the data. Through the completeness evaluation methods, the missing rules can be added into the rule set to ensure good coverage of rules thereby improving the representativeness of the correlation in the data. Evaluating the consistency result strengthens the applicability of the association rules into practice. Combined with the characteristics of the data, the analysis of accuracy, completeness, and consistency indicators can effectively improve the reliability level of the association rules. Furthermore, the general applicability of the proposed reliability evaluation system to the analysis in many research fields is enhanced. Thus, more accurate, complete, and consistent association rules can be achieved.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 41331175. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions which have greatly improved the work.

REFERENCES

Agrawal R,Imielinski T,Swami A, 1993. Mining association rules between sets of items in large databases.In: Proceedings of the ACM SIG-MOD Conference on Management of Data(SI), New York: ACM Press, pp. 207-216.

Berry, J.A., Linoff, G.S, 1997. Data Mining Techniques for Marketing, Sales and Customer Support. John Wiley & Sons,Inc, pp. 45-53.

Dunn J C, 1973. A Fuzzy Relative of the ISODATA Process and Its use in Detecting Compact Well-Separated Clusters. J Cybern, 3(3)，pp. 32-57

Han JW，Cai YD，Cercone N, 1992. Knowledge discovery in database: an attribute-oriented approach . In: Proceedings of International Conference on Very Large Data Bases. (SI): Morgan Kaufmann, pp. 547-559.

Han JW, Kamber M, 2006. Data Mining Concepts and Techniques. Burlington : Morgan Kaufmann.

Hannu T, 1996. Sampling large databases for association rules.In: Proceedings of the 22nd International Conference on Very Large Database, Bombay, India.

Jayakrushna Sahoo, Ashok Kumar Das, A. Goswami, 2015. An efficient approach for mining association rules from high utility itemsets. Expert Systems with Applications, (42), pp. 5754–5778.

Ji GL, Sun ZH, 2004. Mining Optimized Support and Interestingness Quantitative Association Rules. Mini- Micro Systems，25(2), pp. 225-227. (in Chinese)

Katja Windt, Marc-Thorsten Hutt, 2011. Exploring due date reliability in production systems using data mining methods adapted from gene expression analysis. CIRP Annals - Manufacturing Technology , (60), pp. 473–476.

Lou LF, Jiang ZF, Tian SZ, 2003. Studying on the Influence Factor of Interestingness of Association Rules in Data Mining. Computer Engineering and Applications，(6), pp. 190-192. (in Chinese)

Mennis J，Guo D, 2009 .Spatial data mining and geographic knowledge discovery — An introduction[J].Computers ， Environment and Urban Systems, 33(6), pp. 403-408.

Pauray S.M.Tsai, Chen CM, 2004. Mining interesting association rules from customer databases and transaction databases.[J].Information Systems, 29(8), pp. 685–696.

Qi YX, 2004. Measurement of Novelty: Factor of Evaluation for the Association Rules. Application Research of Computers, (1), pp. 17-19. (in Chinese)

Sandro Saitta, Benny Raphael, Ian F.C, 2005. Smith. Data mining techniques for improving the reliability of system identification. Advanced Engineering Informatics , (19), pp. 289–298.

Shi WZ, Chen JP, Xhan QM, Shu H, 2012. Reliable Spatial Analysis. Geomatics and Information Science of Wuhan University，37(8), pp. 883-887. (in Chinese)

Sinha S K. Realiability and lift testing.New York: John Wiley & Sons,1986.

Song XD, Zhai K, Gao WD, 2007. The Research of Evaluation Index in Association Rules. Control & Automation，24(4-3), pp. 174-176. (in Chinese)

Su ZD, You FC , Y ang BR, 2004. Comprehensive evaluation method of association rules and an practical example. Computer Applications，24(10), pp. 17-20. (in Chinese)

Xu F, 2008. Application of the Association Rules Mining Method in the Realibility Information System. Ship Electronic Engineering，28(6), pp. 177-180. (in Chinese)

Zhu WJ, Hu XG, 2007. Survey on Methods of Evaluation for the Association Rules. Journal of Anhui Science and Technology University，21(4), pp. 37-40. (in Chinese)