INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Mining Association Rules Based on Certainty

Liyan Dong[1,2*], **Renbiao Wang**[3], **Yongli Li**[4*]

[1] *College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2] *Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*
[3] *Zhonghuan Information College Tianjin University of Technology, Tianjin 300000, China*
[4] *School of Computer Science, Northeast Normal University, Changchun 130117, China*
[*] *Corresponding authors' Email: dongliyan@gmail.com, liyl603@nenu.edu.cn*

Abstract: The paper proposed a new kind of classification algorithm based on support and certainty, which scanned the same datasets several times to discover certain frequent item sets whose length complied with the fixed increment. The algorithm produced the Boolean association rules by means of the width preference-traversing mode. The experiment shows this algorithm of association rules based on certainty and support architecture could generate a accurate association rules compared with other classification algorithm and improve the accuracy and perceptiveness of association rules effectively.

Keywords: Association rules; certainty; confidence; classification

1. Introduction

In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps.

As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets, which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Apriori uses breadth-first search and a tree structure to count the candidate item sets efficiently. It generates candidate item sets of length k from item sets of length $k-1$. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

Apriori, while is historically significant, suffers from a number of inefficiencies or trade-offs, which have given rise to other algorithms. Candidate set generates large numbers of subsets (the algorithm attempts to load up the candidate set as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset N only after all $N-1$ of its proper subsets.

The Association rules based on confidence and support architecture can be divided into the simple asso-

ciation, the time sequence association and the multidimensional association and so on. Besides, there are some relevant associations among the transaction item sets; what's more, association analysis model has found increasingly wide application in all fields.

Association rules describe certain relationships of the set of attribute items in the database, which is not instructive and simple [1]. Since the classic Apriori algorithm was proposed, the domestic and foreign scholars have focused on the efficiency of the association rules. For instance, Mannila, Toivonen and Verkamo introduced the technology of decision tree pruning to the classic association rules theory [2]. Park, Chen and Yu proposed the association rules algorithm based on hash [3]. Savasere, Omiecinski and Navathe also proposed the association rules algorithm based on partition [4].

Although the previous studies improved the mining efficiency of association rules to some extent, and reduced the space and time complexity greatly of the association rules [5]. However, it has not done much to improve the framework system of the association rules algorithm based on the support and confidence [6]. This paper proposes a new association rule algorithm with framework system based on certainty and support architecture. This algorithm introduces a certainty factor in expert system to measure the certainty of association rules. It accurately describes the relationship between the conditional probability and the prior probability. The algorithm not only resolves the problem of missing association rules fundamentally, but also improves the accuracy of association rules mined.

The introduction of association rule mining in 1993 by Agrawal, Imielinskiand Swami and, in particular, the development of an algorithm by Agrawal and Srikant and by Mannila, Toivonen and Verkamo marked a shift of the focus in the young discipline of data mining onto rules and data bases. Consequently, besides involving the traditional statistical and machine learning community, data mining now attracts researchers with a variety of skills ranging from computer science, mathematics, science, to business and administration. The urgent need for computational tools to extract information from data bases and for manpower to apply these tools has allowed a diverse community to settle in this new area.

The data analysis aspect of data mining is a more exploratory complex search and optimization problem and it is here where mathematical methods can assist most. Particularly, the case for association rule mining requires searching large data bases for complex rules.

Mathematical modeling is required in order to generalize the original techniques used in market basket analysis to a wide variety of applications.

Mathematical analysis provides insights into the performance of the algorithms.

Large amounts of data have been collected routinely in the course of day-to-day management in business, administration, banking, the delivery of social and health services, environmental protection, security and in politics. Such data is primarily used for accounting and for management of the customer base. Typically, management data sets are very large and constantly growing and contain a large number of complex features. While these data sets hide the properties of the managed subjects and relations, and are thus potentially of some use to their owner, they often have relatively low information density. One requires robust, simple and computationally tools to extract information from such data sets. The development and understanding of such tools is the core business of data mining.

2. The Problems Based on Confidence Architecture

Definition1 If the item sets $X \subseteq U$, the total number of transaction of non-empty transaction database is N, the support degree of X is S/N, we denote it as $sup(X)$, that is $P(X)$. If $sup(X) \geq min\_sup$ in which min_sup is the given minimum support threshold, we call X frequent item sets [7].

Definition2 Supposed that $X \Rightarrow Y$ is a association rule, which meets $X \subseteq U$, $Y \subseteq U$, $X \cap Y = \Phi$ the conditional probability of Y is $P(Y|X)$, we call it the confidence degree of $X \Rightarrow Y$, we denote it as $conf(X \Rightarrow Y)$[8].

Definition3 As for the association rules $sup(X \Rightarrow Y) \geq min\_sup$, $conf(X \Rightarrow Y) \geq min\_conf$, we call $X \Rightarrow Y$ as strong association rule [9].

The confidence $X \Rightarrow Y$ is called as conditional probability. Although the confidence degree interpreted by conditional probability reflects the relevance of between X and Y, it separates the relevance of the conditional probability $P(Y|X)$ and the prior probability $P(Y)$, which may lead to three problems as following:

When $sup(X \Rightarrow Y) \geq min\_sup$ and $min\_conf \leq P(Y|X) \leq P(Y)$, the association rule generated is lack of credibility, for $min\_conf \leq P(Y|X)$ output $X \Rightarrow Y$ will be strong association rule. But $P(Y|X) \leq P(Y)$ shows that the probability of Y reduced under the premise of X. Instance X kept Instance Y from advancing. At this

time, it is easy to return the strong association rules $X \Rightarrow Y$ which is normal association rule in fact.

When $sup(X \Rightarrow Y) \geq min\_sup$ and $min\_conf \geq P(Y|X) \geq P(Y)$, we may miss some important association rules. Due to $P(Y|X) \leq min\_conf$, output $X \Rightarrow Y$ would not be strong association rule. But $P(Y|X) \geq P(Y)$ shows that the probability of Y increased under the premise of X. Instance X deduced Instance Y. As a result we ought to get the association rule $X \Rightarrow Y$ as a useful association rule.

When $conf(X \Rightarrow Y) = P(Y|X) = P(Y)$, we can easily get $P(Y,X) = P(X) \times P(Y)$. Instance Y and X are independent. Therefore, the association rules generated was not accurate.

Apriori algorithm only searched k item sets with the same size in terms of fixed increment, and then connected two k-item sets with the last k item sets different so as to generate $k+1$ item sets.

Firstly, the algorithm just searched for the datasets with the size 1 in the database.

Secondly, it compared the datasets with the size 1 with the $min\_support$.

Thirdly, it removed all the datasets with the size shorter than $min\_support$. Thus, we can obtain the frequent item sets with the size 1.

Finally, the algorithm connected two frequent item sets with the size 1 for the purpose of generating the frequent item sets with the size 2. At this time, the algorithm found out the frequent item sets by the formula in order to obtain the enhanced association rules. The reason why the algorithm produced the k-item sets after scanning the database one time is that it can reduce the computational complexity, save time, cut the space cost and improve the algorithm performance.

Apriori algorithm can be divided into two steps during the course of algorithm implementation:

Discover the k-item sets by the means of scanning the same relation database many times and compare the k-item sets with min support previous set so as to obtain the frequent k-item set.

Compare the frequent item sets with the min confidence previous set, and select those frequent item sets with higher confidence to generate the enhanced association rules.

Apriori algorithm usually adopted the method of cooperation between connection and delectation: connection means that it connected two k-item sets with the similar size in order to make up one $k+1$ item sets. First of all, select two k-item sets L_1 and L_2 with the size k. L_{ij} is ranked as the position j of L_i. L_{ik} stood for the position k of L_i.

The purpose of the algorithm is to find out two of those item sets with the first $k-1$ item similar to perform operation. Delectation means that it removed those item sets with support degree smaller than $min\_support$, and the remains of item sets made up $k+1$ frequent item sets.

The pseudo code is as follow:

Input: DataSet, minsup, minconf

Output: ArulesSet Apriori(AruleSet, DataSet, minsup, minconf)

FOR i=1 TO Length(DataSet) do (

FrequentSet[i] $\leftarrow$ GenFrequentSet(i, DataSet)

IF FrequentSet[i] = Φ THEN RETURN.

FOR j=1 TO Length(FrequentSet[i] do)

IF FrequentSet[i][j] $<$ minsup

THEN Remove(FrequentSet[i][j])

AruleSet[i] $\leftarrow$ GenAruleSet(FrequentSet[i], minconf).

DataSet $\leftarrow$ FrequentSet[i].)

Apriori algorithm is a classic data mining algorithm based on width preference traversing model, which produced one-dimensional Boolean Association Rules. The core meaning is to set two important algorithm parameters according to the support and confidence.

Firstly, scan the association database to find out k-item sets.

Secondly, connect two k-item sets to make up $k+1$ item sets.

Thirdly, delete the $k+1$ item sets which failed to meet the $min\_support$ so as to produce the $k+1$ frequent item sets.

Fourthly, derive the association rules expected in terms of $k+1$ frequent item sets.

Finally, remove those association rules not met min confidence.

3. The Theory of Certainty Factor

A certainty factor is used to express how accurate, truthful, or reliable you judge a predicate to be. It is your judgment of how good your evidence is. The issue is how to combine various judgments.

Note that a certainty factor is neither a probability nor a truth value.

Consider the expression George is suffering from hypoxia.

Based on warnings given to pilots, we would speak of there being strongly suggestive evidence that George is suffering from hypoxia when he is flying in an unpressurized airplane at 4,000 meters (13,000 ft) and

his judgment, memory, alertness, and coordination are off.

Note, we are not saying "there is an eighty percent chance that George suffers hypoxia"; that is a probability estimate. We are talking about our judgment of certainty. You may be able to generate statements of probability, such as: "80% of US Air Force student pilots will fail to maintain altitude within 100 feet when they fly higher than ... meters without supplementary oxygen, and this will indicate they suffer from hypoxia." But this is a different sort of statement involving certainty factors.

What I am doing in this example of uncertainty is taking what I was taught as a student pilot and creating from that information a mechanism for diagnosing hypoxia. I don't know the probability that a person of my health and age will suffer hypoxia at 4,000 meters but I do know the symptoms, which, however, may be weak, or have other causes.

In McAllister's scheme, a certainty factor is a number from 0.0 to 1.0. A phrase such as suggestive evidence is given a number such as 0.6; strongly suggestive evidence is given a number such as 0.8. The person making the judgment uses the scale more or less as an ordinal scale. The numbers are used in a metric to permit a computer to make calculations.

McAllister's rules for combining certainty factors are such that you can add new evidence to existing evidence. If the evidence is positive, this increases your certainty, as you would expect. But you never become 100%.

Continuing our hypoxia example: George tells us that he feels wonderful. This is suggestive evidence that George suffers from hypoxia. (Pilots are warned of this: "if you feel euphoric, consider hypoxia: you may be flying too high without oxygen, or suffering carbon monoxide poisoning from a broken heater." Of course, there are many good reasons to become euphoric when you fly; hypoxia is insidiously dangerous.)

The association rules based on the confidence and support can be divided into the simple association, the time sequence association and many dimensions association.

Besides, association analysis model was used widely in the mutual relation domain which happened among the transaction item sets.

Although the Apriori algorithm based on support and confidence architecture has possessed of complete theory basis and was applied to every walk of life, however, the support and confidence were set depending on personal experience completely, besides, it lacked related data and international general standard, accordingly, if we want to apply the algorithm to the particular company CRM (Customer Relationship Management) system, it may exit some instability and defect.

There're some defects in Apriori association rules algorithm based on support and confidence architecture. The reason why it has some defects can be summed up as follow:

$$conf() = \frac{\sum_{(A\cap B)\in I_j} I_i}{\sum_{A\in I_j} I_j} = \frac{P(A\cap B)}{P(A)} = P(B|A) \qquad (1)$$

where $P(A\cap B) = P(AB) = P(A)\cdot P(B)$,
$P(B|A) = P(B)$, $P(B) > Min\_conf$.

From what was discussed above, we can draw the conclusion that there's no relationship between the prior part and back part of association rules. If only the probability in which back part of association rules happened is larger than min confidence, the association rules can be returned to user as enhanced association rules. However, in fact, this association rule made no sense, moreover, it's misleading.

It's because there exists the deadly defects in the association rules based on support and confidence architecture that this paper proposed new association rules architecture-support and certainty architecture.

The theory of certain factor is proposed by Shortliffe in 1975, which was a reasoning model-MYCIN. The reasoning model depended on the certainty factors theory, which was proposed aimed at Bayes conditional probability.

The theory was used to solve the uncertainty problem and was applied to expert system based on medical diagnosis successfully.

As is known to all, applying Bayes to the medical diagnosis, we need to know some parameters, such as prior probability. However, it's difficult a lot.

For example, there're many production lines in a certain manufacturing workshop. Now we want to test the conditional probability of which it came from number i production line in the condition which it's unqualified. Bi represents the product from i production line, A represents the product which is unqualified, the Bayes conditional probability formula is as follow:

$$P(B_i|A) = \frac{P(A|B_i)\cdot P(B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)} \qquad (2)$$

Generally speaking, it's so difficult to obtain the probability knowledge. What's more, with the evidence increased, it needs more previous knowledge.

Traditional probability emphasizes on the sum between the probabilities in which certain hypothesis happened and the probabilities in which the same hypothesis didn't happen equaled to 1.

$$P(H/E)+P(\neg H/E)=1$$

In fact, the possibility in which hypothesis happened depending on certain evidence of back experience equaled to 0.7. However, hypothesis proposed depending on certain complementary events of evidence of back experience may not equal to 0.3.

E1: creature chromosome is gram-positive.

E2: creature structure is ball germ.

E3: creature shape is chain appearance.

H: reference evidence with intensity 0.7 can prove that the creature belongs to hammer-throwing bacterium.

$$P(H/E_1 \cap E_2 \cap E_3)=0.7$$

The formula comes from the knowledge of internal medicine expert in MYCIN knowledge library, which doesn't comply with the expert's opinion.

In fact, the question is a trusty tolerance value which internal medicine expert provided depending on the evidence introduced above. However, the evidence above which was deduced to testify that it wasn't hammer-throwing bacterium which doesn't exist any trusty tolerance. Accordingly, the method that we expressed such association rules ainappropriate.

The specific formulas are as follows:

$$MB(H,E)=\begin{cases} 1 & P(H)=MAX[0,1] \\ \frac{MAX[P(H/E),P(H)]-P(H)}{MAX[1,0]-P(H)} \end{cases} \quad (3)$$

$$MD(H,E)=\begin{cases} 1 & P(H)=MIN[0,1] \\ \frac{MIN[P(H/E),P(H)]-P(H)}{MIN[1,0]-P(H)} \end{cases} \quad (4)$$

$$CF(H,E)=MB(H,E)-MD(H,E) \quad (5)$$

CF is the certainty factor based on some assumption H under the situation that the posterior evidence exists. $MB(H,E)$ is the increased confidence measure caused by the existence of E. $MD(H,E)$ is the increased no-confidence measure caused by the existence of E. The relationship between $MB(H,E)$ and $MD(H,E)$ can be expressed by table 1.

Figure1 illustrates the experiment relation between $CF(H/E)$ and $P(H/E)$. When the condition probability is larger than 0, the certainty is positive value.

Table 1 Definition of terms

conclusion	MB, MD, CF
If $P(H/E)=1$ is true	$MB=1, MD=0, CF=1$
If $P(\neg H/E)=1$ is false	$MB=0, MD=1, CF=-1$
evidence	$MB=0, MD=0, CF=0$

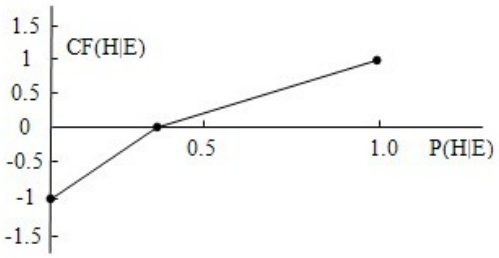

Figure 1 Experiment Relation Figure

MB and MD met the mutual exclusion relation, because for the same back evidence there couldn't be opposition and agreement:

$MB(H/E)>0, MD(H/E)=0$

$MD(H/E)>0, MB(H/E)=0$

$CF>0$, $MB>MD$, evidence support hypothesis.

$CF>0$, $MB=MD=0$, there's no relation

If $CF(Y,X)>0$, it shows the appearance of X increases the probability of Y. The larger the value of $CF(Y,X)$ is, the greater the increase probability Y is.

If $CF(Y,X)<0$, it shows the appearance of X decreases the probability of Y. The smaller the value of $CF(Y,X)$ is, the greater the reduction probability Y is.

Thus the association rules based on support and certainty architecture described the mutual relationship among the dataset accurately compared with the association rules based on support and confidence architecture, namely, the trust degree of prior part of association rules relative to back part of association rules. So far as probability theory is concerned, the situation in which some events happened hindered other situation in which the next events happened.

As above, the paper introduced certainty factor into architecture of association rules. When one association rule met *min_support* and certainty, it was called as certainty association rules. Under the architecture based on support and certainty, the association rule complied with classification data mining algorithm was proved to be certainty association rules. As it were, when a piece of association rule is larger than the *min_support* and meets the condition of certainty, it can be returned to the real planner as one valuable association rule.

4. Algorithm Description

In general, we should set certain threshold to fix on the conditions in which association rules met in advance, the threshold $\Theta \in (0,1]$, the range of Θ based mainly on the association strength that the classification system required. When $\Theta = 1$, the evidence of posterior supports the hypothesis proposed fully. This paper takes $\Theta > 0$.

Scan all k-item sets in the database, find all k-item sets whose support is greater than the minimum support by connecting operation, deleting operation, etc. Then generate k-frequent sets.

Calculate the certain degree of k-frequent sets and find all k-frequent sets that meet the certain condition and derive the certain association rules.

The classification technology based on association rules was used to mine practical, valuable, constructive, no error rules from massive, ruleless, no comprehensive, noisy, rough, uncertain dataset in substance.

Apriori algorithm based on support and confidence architecture sets the parameters depending on the personal experience and lacked of relative statistics. Besides, it couldn't describe the relation between support and confidence. This paper implemented a new kind of classification algorithm based on support and certainty architecture—- CBCFA(Classification Based on Certainty Frame Apriori).

The most important characteristic of this method is that it adopts the architecture of certainty and supports instead of the architecture of confident and support. In that certainty theory can reflect the relationship among the datasets more accurately, namely, the trusty degree in which the prior part of association rules deduced the back part of association rules. Accordingly, it has profound instructive meaning for this problem.

CBCFA Algorithm based on certainty and support architecture can be interpreted as follow:

For example, the initial datasets is a relation database, which stored the relative transactions. The relation database included several tables, which contained l general attributes and k classification determinant attributes. Besides, there're n piece of data records (n lines of relation tuples). According to the category attribute they can be classified in k known categories. Furthermore, we need to do the data preparation, which incorporates data integration, data selection, data preprocessing.

Data integration:

It mainly imports several documents or several relation tables of transaction database into one dataset. We can perform unified data processing, clear the noisy data and do the standardization by means of OLAP (Online Analytical Processing).

Data selection:

It mainly performed analysis, abstraction for the integrated datasets, selected data objects which data mining algorithm needed and shrunk the data scales in order to live up to decrease the cost and performing time in order to improve the algorithm performance. For example, for a sheet of relation data table, we just select the valuable, profound instructive meaning to analyze and discard useless information.

Data preprocessing:

In that different data mining algorithm adopted different mining processes. Accordingly, its scope of application is different.

Some algorithms are sensitive to noisy data, however others are not, such as rough set. Some algorithms need a lot of knowledge with previous experience, such as probability theory algorithm.

The most classification algorithms are sensitive to input of continuous attributes and need to perform discretization treatment.

For instance, the algorithm which the paper implementation needs to go through data integration, data selection above mentioned. Besides in the data preprocessing we need to perform discretization treatment on general attribute of relation database in order to sort out data item sets in the form of standardization more reasonably.

Discretization: In mathematics discretization concerns the process of transferring continuous models and equations into discrete counterparts. This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers.

In order to be processed on a digital computer another process named quantization is essential.

CBCFA algorithm obtained the association rules met the $min\_support$ and certainty by means of scanning the data in the transaction database. Classification-SC algorithm description:

Input: data item sets k, minimum support threshold min_sup, minimum confidence threshold min_conf.

Output: all the item sets that meet the requirements.

Step 1 Initialize the database and scan all the item sets that the length of them is k according to the initial conditions, and the initial value of k is 1.

Step 2 Calculate the support of each k-item set, remove the k-item set whose support is less than min_sup. If the k-item set is empty, we finish the algorithm.

Step 3 Connect the k-item set in accordance with

Apriori. If the $k+1$-item set are empty, we finish the algorithm.

Step 4 Calculate the support of each $k+1$ item set, remove the k-item set whose support is less than *min_sup*. If the k-item set is empty, we finish the algorithm.

Step 5 Calculate the certain of each $k+1$-item set, remove all $k+1$-item set whose certainty is less than threshold Θ. If $k+1$-item set are empty, we finish the algorithm.

Step 6 Get the certain association rules according to the result of $k+1$-item certain frequent set, then $k++$.

We use pseudo-language to describe the classification algorithm based on certain-support framework.

Input: DataSet, minsup

Output: AruleSet // the association rules set of certain

classification-SC(AruleSet, DataSet, minsup, cer_threshold)

Step 1 FOR i=1 TO Length(DataSet) do

Step 2 FrequentSet[i] ← GenFrequentSet(i,DataSet)

Step 3 IF FrequentSet[i] = THEN RETURN //if i-item data set is empty, finish the algorithm

Step 4 FOR j=1 TO Length(FrequentSet[i]) do

Step 5 IF FrequentSet[i][j] $<$ minsup THEN Remove(FrequentSet[i][j])

Step 6 AruleSet[i] ← GenAruleSet(FrequentSet[i], cer_threshold) //generate *i*-item certain association rules

Step 7 DataSet ← (FrequentSet[i]) // the $i+1$-item frequent set is got by *i*-item frequent set.

The most important step of CBCFA is second phase-producing rules. The input value is general rule sets and the output is certainty rule sets. Then we put the rule sets in the rule library according to categorizer construction algorithm.

The algorithm which the paper proposed need to determine two parameters— minsup and Θ, and the setting of *min_support* can refer to Apriori algorithm, $\Theta \in (0,1]$ must be appropriate. If the setting is too small, then the trusty degree of evidence relative to the rule is lacking. If the setting is too big, then k item sets which met the certainty are not enough. Therefore, it should make the analysis of concrete conditions.

This paper proposed a new kind of classification algorithm based on support and certainty, which scanned the same datasets several times to discover certain frequent item sets whose length complied with fixed increment. The algorithm produced the Boolean association rules by means of the width preference traversing mode. The next flow chart is the execution procedure of the algorithm:

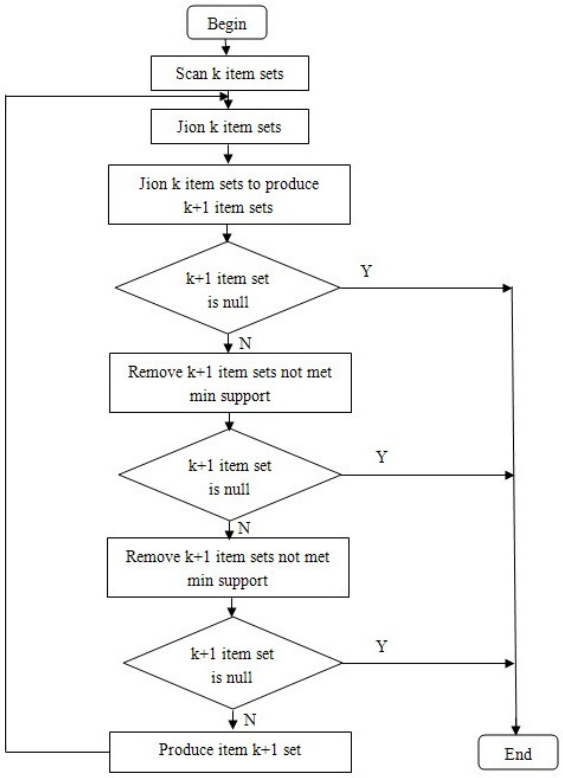

Figure 2 Simulated experiment

Table 2 Trading statistics

	pens bought	pens not bought	total
The pencils bought	4200	2800	7000
The pencils not bought	1800	1200	3000
Total	6000	4000	10000

5. Simulated Experiment

Example 1: The value of *min_sup* is 30% and the value of *min_conf* is 60%.

The analysis of Apriori algorithm is as follows:

$Q\ sup = \frac{4200}{10000} = 42\% > min\_sup$

$conf = \frac{4200}{6000} = 70\% > min\_conf$

$\therefore\ buy(pen) \Rightarrow buy(pencil)$ is a strong association rules.

The analysis of classification-SC algorithm is as follows:

$sup = \frac{4200}{6000} = 70\% > min\_conf$

According to the certain formula:

$CF(H,E) = MB(H,E) - MD(H,E)$

$\because\ 0 < P(B) < 1$

$\therefore\ CF(B,A) = MB(B,A) - MD(B,A) = 0$

$\therefore\ CF(B,A) < \Theta$

$\therefore\ buy(pen) \Rightarrow buy(pencil)$ isn't a strong association rules.

Table 3 The relationship among the product parts

	Defective products	Qualified products	total
equipment from B	52	636	688
equipment from others	11	1800	1811
Total	63	2436	2499

The percentage of customers who have bought pencil is 70% in all the customers who have bought pen. The average probability of customers bought pencil in all customers is the same with the conditional probability of customers who have first bought pen and then have bought pencil. Therefore, we can see from the above analysis that the behavior of buying pen did not promote the behavior of buying pencil.

Example 2 The value of min_sup is 20% and the value of min_conf is 20%.

The analysis of Apriori algorithm is as follows:

$\because sup = \frac{52}{2499} = 2.08\% > min\_sup$

$conf = \frac{52}{688} = 7.56\% < min\_conf$

$\therefore produce\ (equipment - from - B) \Rightarrow result\ (defective - products)$

The association rule isn't a strong association rule.

The analysis of classification-SC algorithm is as follows:

$sup = \frac{52}{2499} = 2.08\% > min\_sup$

$\because 0 < P(B) < 1$

$\therefore MB(B,A) = \frac{P(B/A)-P(B)}{1-P(B)} = \frac{0.07558-0.02521}{0.9748}$
$= 5.2\% > \Theta$

$\therefore MD(B,A) = 0$

6. Conclusions

The paper proposed a new kind of classification algorithm based on support and certainty, which scanned the same datasets several times to discover certain frequent item sets whose length complied with fixed increment. The algorithm produced the Boolean association rules by means of the width preference traversing mode. The experiment shows this algorithm of association rules based on certainty and support architecture could generate association more accurate rules compared with other classification algorithm and improve the accuracy and perceptiveness of association rules effectively.

In conclusion, the paper introduced certainty factor into architecture of association rules. When one association rule met $min\_support$ and certainty, it was called as certainty association rules. Under the architecture based on support and certainty, the association rule complied with classification data mining algorithm was proved to be certainty association rules. As it was, when a piece of association rule is larger than the $min\_support$ and meets the condition of certainty, it can be returned to the real planner as one valuable association rules.

The strong association rules got from the classic association rules algorithm based on support-confident framework might go against the objective laws of reality. This is because the association rules based on confident-framework cannot describe the relationship between the evidences and the assumptions accurately. In addition, we cannot measure the influence which prior event had on the back event. The paper introduced certainty factors to verify the confidence degree, accordingly describing the relationships between the conditional probability and prior probability. What's more, the algorithm resolved the problem that important association rules often were missed.

References

[1] Zhun Zhou, Bingru Yang, Yunfeng Zhao, Wei Hou, "Research on algorithms for association rules mining based on FP-tree", *Systems and Control in Aerospace and Astronautics*, ISSCAA 2008, 2nd International Symposium on, pp.1-5.

[2] R.Agrawal, T.Imielinski and A.Swami, "Mining association rules between sets of items in large databases", *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, ACM Press, pp.207-216, 2003.

[3] Heikki Mannila, Hannu Toivonen and A. Inkeri Verkamo. "Efficient Algorithms for Discovering Association Rules", *Proceedings of AAAI'94 Workshop on Knowledge Discovery in Databases*, Seattle, Washington. AAAI Press, pp.181-192, 1994.

[4] PMehmet Kaya, PReda Alhajj, "Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining", *Applied Intelligence*, vol.24, pp.7-15, Feb. 2006.

[5] Amihood Amir, Yonatan Aumann, Ronen Fildman, Moshe Fresko, "Maximal Association Rules: A Tool for Mining Associations in Text", Journal of Intelligent Information System, vol.25, pp.333-345, Nov. 2005.

[6] Davy Janssens, Geert Wets, Tom Brijs, Koen Vanhoof, "Adapting the CBA algorithm by means of intensity of implication", *Information Sciences*, vol.173, pp.305-318, June 2005.

[7] Muhammad Fauzi Mohd. Zain, Md. Nazrul Islam, Ir. Hassan Basri, "An expert system for mix design of high performance concrete", *Advances in Engineering Software*, vol.36, pp.325-337, May. 2005.

[8] Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen, "A comparison of implicit and explicit links for web page classification", *Proceedings of the 15th international conference on World Wide Web*, pp.643-650, May. 2006.