Improved Data Mining Algorithms Based on an Early Warning System of College Students

Lijuan Zhou
Department of Information Engineering College, Capital Normal University, Beijing, China
zlj87@tom.com

Yuyan Chen and Shuang Li
Capital Normal University, Information Engineering College, Beijing, 100048, China
cyy5112360@163.com, lishuang924@163.com

Abstract—**In order to solve the problem of early warning of college students' achievement, this paper proposes two improved algorithms for data pre-processing and mining warning factor. At first we put forward an improved K-Means algorithm which is not only ensures the accuracy of the original algorithm, but also improves the stability of the algorithm. Then we put forward an improved algorithm New_Apriori algorithm and analyze experiment result. The result shows that the amount of data access has been reduced significantly and efficiency has been improved. In the end of this paper, we built the early warning model of students' achievement based on neural network. The result of experiment shows that the new algorithms improve the efficiency and accuracy of the early warning.**

Index Terms—**Data Mining; Cluster Analysis; Association Rules; BP Neural Network**

I. Introduction

In recent years, higher education develops rapidly. However, the amount of college student who was discouraged or suspended due to lack of sufficient course credits increases quickly. Meanwhile, with the continuous progress of educational informationization and the development of database technology, education data has been increasing quickly [1]. We can mine potential, reliable and credible information from these a lot of educational information. This provides guidance information for research on early warning [2, 3] about students' grade. In addition, data mining technology [4, 5] develops rapidly, so data mining technology can be applied to research on early warning about students' grade and achieves the purpose of early warning effectively, comprehensively and objectively.

Frequent itemset mining is a core data mining operation and has been extensively studied over the last decade [6]. Algorithms for frequent itemset mining form the basis for algorithms for a number of other mining problems, including association mining, correlations mining, and mining sequential and emerging patterns. Algorithms for frequent itemset mining have typically been developed for datasets stored in persistent storage and involve two or more passes over the dataset [7]. Recently, there has been much interest in reducing times of scanning and stored memory to improve algorithms.

In order to solve the problem of early warning of college students' achievement, we propose two improved algorithms for data pre-processing and mining early warning factor. Firstly we propose an improved K-Means algorithm for data pre-processing. In K-Means algorithm, the initial cluster center is random distribution. That causes fluctuation effects on cluster result. The improved algorithm not only ensures the accuracy of the original algorithm, but also improves the stability. Then we propose an improved New_Apriori algorithm and analyze experiment result. Experiment result shows in the new algorithm, the amount of data access has been reduced significantly and efficiency has been improved. The Algorithm especially suits finding the largest number of frequent itemsets only. In the end of the paper, we build a BP neural model to mine warning information. The result of experiment shows that the new algorithms improve the efficiency and accuracy of the early warning.

II. Use Optimized K-Means Algorithm to Seek Initial Clustering Center

A. K-Means Clustering Algorithm

There are four common clustering algorithms: partitioning algorithm, hierarchical algorithms, large database clustering and clustering to classification attribute [8]. Among these algorithms, in this paper we use one of the most common partitioning algorithms: k-means algorithm, mainly because k-means algorithm is a classical algorithm to solve clustering problems. It's simple, fast and it can deal with large data efficiently. Therefore, we choose k-means algorithm to make clustering analysis for students' grade data.

K-means algorithm was put forward by J.B.MacQueen in 1967[9]. It's the most classical

doi:10.4304/jsw.8.9.2352-2359

clustering algorithm that has been widely used in science, industry and many other areas, which has produced deep influence. K-means algorithm belongs to the partitioning algorithm. It's an iterative clustering algorithm. In the iterative process, it keeps moving the members of the cluster set till we get the ideal cluster set. The members of the cluster are highly similar. At the same time, the members of different clusters are highly diverse. Ki= {ti1, ti2… tim}, define its average as $m_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij}$. K-means algorithm needs the number of expected clusters to serve as parameters input. Its core idea is: input the number of expected cluster: K, divide N tuples into K clusters. It makes the members of the cluster are highly similar and the members of different clusters are highly diverse. The cluster average that is given above is the cluster centroid. So we can calculate the similar degree or the distance between clusters according to the cluster centroid.

K initial cluster centers of k-means algorithm are allocated randomly or use the previous k objects directly. Different initial cluster centers lead to different clustering results and the accuracy will also change. And using clustering algorithm to thin the classification of students' grade is the primary step of the students' grade early warning research. It has deep influence to the future research. In view of this, aiming at fixing problem of initial cluster centers, we put forward an optimized k-means algorithm in this paper.

B. Seek Initial Clustering Center using Optimized K-Means Clustering Algorithm

Firstly, in response to K-Means algorithm's flaw, combining with the practical application of early warning about students' grade, an optimized algorithm is proposed for data pre-processing. In K-Means algorithm, the initial cluster center is random distribution. That causes fluctuation effects on cluster result. The improved algorithm not only ensures the accuracy of the original algorithm, but also improves the stability of the algorithm.

There are two important concepts in optimization algorithm:

Density parameter: Centering object ti, constant Pts objects are contained in radius r. Then r is called the density parameter of ti, we use Ri to represent it. The bigger Ri is, the lower the density is. Otherwise, it means that the regional data density is higher.

Unusual point: This is obviously different from other objects in data set.

Optimized K-Means algorithm initial cluster centers selecting algorithm:

1) Calculate each object's density parameter Ri of set S to compose a set R. (All objects compose set S)

2) Find minimum Rmin of set R, that is, in the region where the object is, the data density is the highest. Treat this object as a new initial cluster center. Delete this cluster center and Pts objects in its range from S. Delete all density parameters from R.3) Repeat step 1), 2) until finding out K initial cluster centers.

The flow chart of the optimized selecting initial cluster center algorithm is fig. 1.

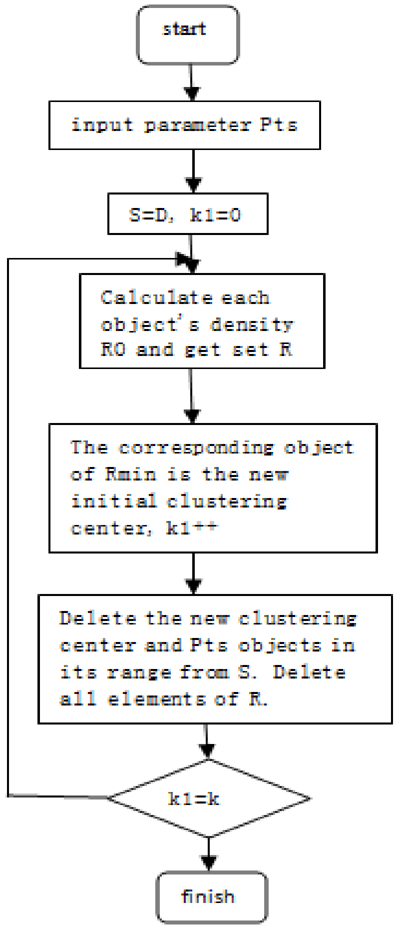

Figure 1. The Flow Chart of the Optimized Selecting Initial Cluster Centers Algorithm

As can be seen, setting constants Pts is the most essential thing in the optimized K-Means algorithm for the initial cluster centers. The range of density parameters may include the unusual point if Pts is large. This will inevitably affect the final result. On the contrary, if Pts is small, the k initial cluster centers may be too concentrated to response the distribution of data initially. After repeated experiments, the ideal range of Pts is [N/k-5, N/k-1]. If N is large and K is small, Pts tends to N/k-5 is better. On the contrary, Pts is better to tend to in favor of N/k-1. After determining the k initial cluster centers, we get the clustering results through applying the K-Means algorithm from the beginning of k cluster centers.

In order to examine the effectiveness of the initial cluster centers selecting optimized algorithm, we experimented with students' grade. Data sources are from all students in 4 different grades in a normal school. In our experiment, we divide the number of clustering cluster into 2 kinds, one is 4 and the other one is 5.We analyze the results that have 4 clusters in detail and divided students into four clusters as cluster0, cluster1, cluster2 and cluster3. Cluster2 are the students with lower scores that we pay close attention to. Students of cluster0 and cluster1 have a good achievement and the cluster3 is in the middle. At last we find out that the result of four clusters is more accurate when comparing example of clusters with the reality. The experiment shows that the optimized algorithm makes up for deficiencies of the algorithm and improves the stability of the algorithm as well as ensures the accuracy of the original algorithm.

III. Basic Idea And Application of New_Apriori Algorrithm

A. *Basic Concept of Association Rules*

Set I = (i1, i2... im) being the set of item and D for the set of transaction. Here transaction T is a collection of items, and $T \subseteq I$. A unique identifier corresponds to each transaction, such as transaction number, denoted by TID. Assume that X is a collection of items, if the $X \subseteq T$, so that transaction T contains X. An association rule is the implication of the form $X \Rightarrow Y$, where $X \subset I, Y \subset I$, and $X \cap Y = Q$. The degree of support of rule $X \Rightarrow Y$ in the transaction database D is a ratio of the number of transaction contains both X and Y and the number of all transactions, denoted by support($X \Rightarrow Y$), that is:

$$\text{Support}(X \Rightarrow Y) = \{T: X \cup Y \Rightarrow T, T \in D\}/D \times 100\% = S\%$$

The confidence of rule $X \Rightarrow Y$ in the transaction set D is the ratio of the number of transaction that contains the X and Y and the number of transaction that contains X, denoted by confidence ($X \Rightarrow Y$), which is:

$$\text{Confidence}(X \Rightarrow Y) = \{T: X \cup Y \subseteq T, T \in D\}/T: X \subseteq T, T \in D\} \times 100\% = C\%$$

Confidence can also be expressed by support, as follows:

$$\text{Confidence}(X \Rightarrow Y) = \text{Support}(X \Rightarrow Y)/\text{Support}(X) \times 100\% = c\%$$

Frequent itemset means the itemset whose support is not less than the minimum support.

Given a transaction set D, the problem of mining association rule is to generate the association rule, whose support and confidence are greater than the minimum support(min_sup) and minimum confidence(min_conf) separately given by the user. In order to mine the meaningful association rule, two thresholds are needed to set: minimum support and minimum confidence. The former means that the minimum level of a set of item in statistical sense. The latter reacts the minimum reliability of association rule [10, 11].

B. *Basic Concept and Application of New_apriori Algorithm*

Against the complexity of early warning factor about students' grade and the original algorithm's, we puts forward an improved algorithm New_Apriori algorithm and analyze experiment result after the data pre-processing. Experiment result shows in the new algorithm, the amount of data access has been reduced significantly and efficiency has been improved. The Algorithm especially suits finding the largest number of frequent itemsets only.

Literature [12] introduces two pieces of inference as follows:

Inference 1: If the count of item in K-frequent item set Lk is smaller than or equal to k, then the item set is the sets of the largest frequent item set.

Inference 2: The count of item in the largest frequent item set Lk is smaller than or equal to the largest count k of items, which satisfies support in all of transactions.

The New_Apriori algorithm makes use of above-mentioned inference, at first making sure the count k of item of the largest frequent item set. It directly gets the largest frequent item set Lk among these transactions, whose count of item is larger than or equal to k. Then consider these transactions one by one in order, whose count of item is k-1 and k-2.Dine the frequent item set, which does not belong to the largest frequent item set with the largest count of item.

As follows introduce the New_Apriori algorithm with an example:

Table 1 is an original affair database. Each affair T represents a student's record; the I1-I7 means different attributes respectively. Minimum support min_support =4.

TABLE I.
ORIGINAL AFFAIR DATABASE

Transaction	Itemset	Flag1	Transaction	Itemset	Flag1
T_1	I_1,I_7	2	T_{11}	I_4,I_7	2
T_2	I_2,I_3,I_4	3	T_{12}	I_1,I_2,I_3,I_4,I_5	5
T_3	I_4,I_5,I_7	3	T_{13}	I_2,I_4,I_5,I_6,I_7	5
T_4	I_2,I_4,I_5	3	T_{14}	I_2,I_3,I_4	3
T_5	I_5,I_6,I_7	3	T_{15}	I_4,I_5,I_7	3
T_6	I_2,I_4	2	T_{16}	I_3,I_4	2
T_7	I_3,I_5,I_7	3	T_{17}	I_1,I_3,I_5,I_7	4
T_8	I_2,I_4,I_5,I_6	4	T_{18}	I_2,I_4,I_5,I_6	4
T_9	I_5,I_7	2	T_{19}	I_5	1
T_{10}	I_2,I_4,I_5	3	T_{20}	I_2,I_4,I_5,I_6	4

Scan the database to take down C[n], C[1]=1, C[2]=5, C[3]=8, C[4]=4, C[5]=2;and the Flag1mark of each transaction. At the same time, change data storage structure, taking item as the key word:

TABLE 2
DATA STORAGE TAKING ITEM AS THE KEY WORD

Item	Transaction	Flag2
I_1	T_1,T_{12},T_{17}	3
I_2	$T_2,T_4,T_6,T_8,T_{10},T_{12},T_{13},T_{14},T_{18},T_{20}$	10
I_3	$T_2,T_7,T_{12},T_{14},T_{16},T_{17}$	6
I_4	$T_2,T_3,T_4,T_6,T_8,T_{10},T_{11},T_{12},T_{13},T_{14},T_{15},T_{16},T_{18},T_{20}$	14
I_5	$T_3,T_4,T_5,T_7,T_8,T_9,T_{10},T_{12},T_{13},T_{15},T_{17},T_{18},T_{19},T_{20}$	14
I_6	$T_5,T_8,T_{13},T_{18},T_{20}$	5
I_7	$T_1,T_3,T_5,T_7,T_9,T_{11},T_{13},T_{15},T_{17}$	9

Because C [5] <min_support, C [4]>min_support, the largest count k of item of the largest frequent assumed before is 4. Now from the transformed data storage select the item with Flag2≥min_support. Then in its transaction set we look into the transaction whether its Flag1i is larger than k. If such transaction's count satisfies the minimum support, take down temporarily. Then get the lately small database (table 3):

TABLE 3
THE LATELY SMALL DATABASE

Item	Transaction
I_2	$T_8,T_{12},T_{13},T_{18},T_{20}$ (U_2)
I_4	$T_8,T_{12},T_{13},T_{18},T_{20}$ (U_4)
I_5	$T_8,T_{12},T_{13},T_{17},T_{18},T_{20}$ (U_5)
I_6	T_8,T_{13},T_{18},T_{20} (U_6)

Now try to get U2 ∩ U4 ∩ U5 ∩ U6. If the count of item in the intersection is larger than or equal to min_support, then {I2,I4,I5,I6} is the largest frequent itemset according to the inference in literature[13]. We get the intersection {T8,T13,T18,T20}, and the count of item is 4. This satisfies the minimum support min_support. Thus it can be seen, {I2,I4,I5,I6} is the largest frequent itemset.

Owing to there are just I2,I4,I5,I6 four items in this experiment, it is equal to the largest count k of item of the largest frequent assumed. So in the process of solving, when we try to get U2∩U4=Ua, if the count of Ua's item is smaller than 4, |Ua|<4, there is no need continue to get the intersection. Because the count doesn't satisfy the minimum support, at this time the count of considering item is smaller than k. Then there isn't the largest frequent itemset. If we don't find 4-frequent itemset, then the largest count k of item of the largest frequent assumed decrease 1.Get the largest frequent itemset with the method above-mentioned.

According to the specialty of Apriori algorithm: All of the son set of frequent itemset are frequent itemset, we can from the largest frequent itemset{ I2,I4,I5,I6} get part of 2-frequent itemset{ I2,I4},{ I2,I5},{ I2,I6},{ I4,I5},{ I4,I6},{ I5,I6} and part of 3-frequent itemset { I2,I4,I5}etc.

If we are interested in other 2-frequent itemset and 3-frequent itemset portably existing in the database, we can continue to try to get the largest frequent itemset to database. We just need to take the k as 2 or 3. In this experiment , at last we find all of the largest frequent itemset{ I2,I4,I5,I6},{ I3,I4},{ I4,I7},{ I5,I7}. And {I2,I4,I5,I6} is the largest frequent itemset with the largest count of item.

The flow chart of the New_Apriori algorithm is fig.2.

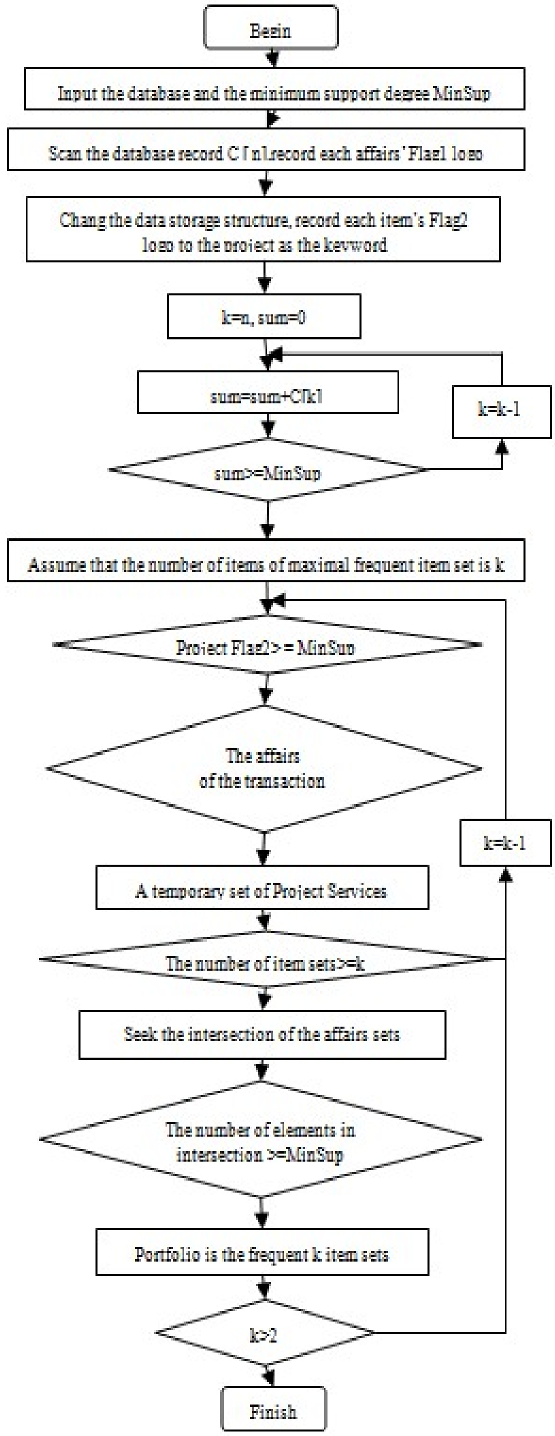

Figure 2. The Flow Chart of New_Apriori Algorithm

This algorithm no longer visits the whole databases when dine the largest frequent itemset with the largest count of item. It only concerns the transaction, whose count of item is larger than or equal to the largest count k of item of the largest frequent assumed. At the same time, the data is no longer saved with the affair ID as the key word. We take the item ID as the new key word. We can save storage space for the database with huge amount of transactions. And particularly there is non-intelligence factor in the student's early-warning database, existing sparse data. The algorithm no longer interview database to compute the support. However, it tries to get the intersection of items' transaction sets, simplifying the process. After dining the largest frequent itemset with the largest count of item, if we are interested in other son frequent itemset, which doesn't belong to it, we can do the same process on the spare data.

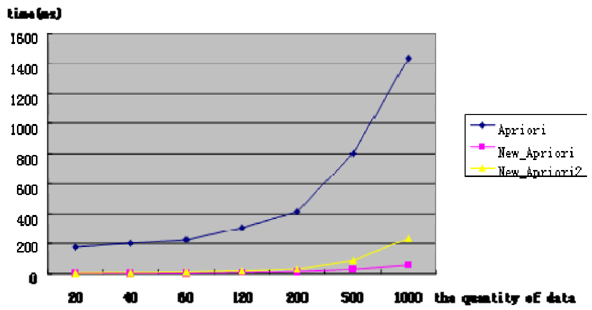

Figure 3. The Result of the Experiment

Fig.3 is the experiment result. The x-coordinate is the quantity of data, and the y-coordinate represents consuming time. The rhombic folding line marks the time of classic Apriori algorithm consuming along with the quantity of data. The quadrate folding line means the consuming time when the new algorithm only dines the largest frequent itemset with the largest count of item. And the triangular folding line represents the consuming time of the new algorithm dines all of frequent itemset. We can see:

The new algorithm's efficiency rises obviously.

Along with data quantity aggrandizement, Apriori algorithm's consuming time changes more obvious than the new algorithm. Apriori algorithm's efficiency lowers greatly.

Along with data quantity aggrandizement, the consuming time seldom changes when we only get the largest frequent itemset with the largest count of item.

We discretized on students' achievement according to the characteristics of algorithm data type and mined the students' achievement by using the New_Apriori Algorithm. We analyzed the experimental results, and prepared for the next BP neural network warning.

IV Built The Early Warning Model Of Students' Achievement Based On Bp Netral Network

Basing on the early warning factor about students' grade, we build BP neural model [14, 15] to mine warning information.

A. Built Students' Grade early-warning Model

We mine several association rules from the poor performance students' grade that we already have. The rules related with all the courses that students learned during 4 years. In these courses, program comprehensive design is the last one to learn. So now we try to use other courses' grades to predict its grade.

BP neural network data comes from all students in 4 different grades in a normal school. We use one class of students' data to predict and the others to train the network.

First we normalize all course grades, that is to say normalize them from 0 to 1.

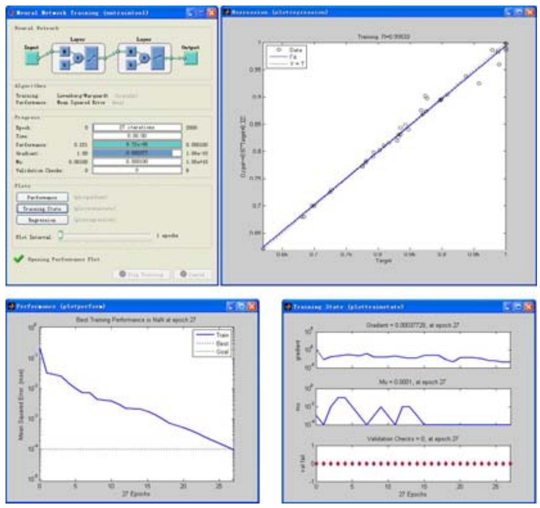

Figure 4. Neural Network Training Results

We obtain the deviation simulation results and show it in figure 5:

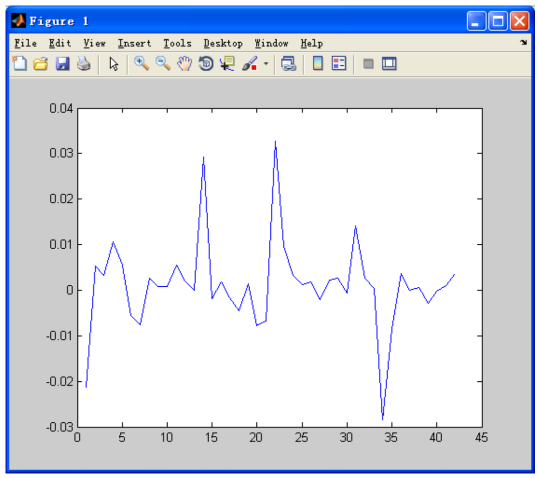

Figure 5. BP Neural Network Deviation Simulation Results

We obtain the deviation testing results and show it in figure 6:

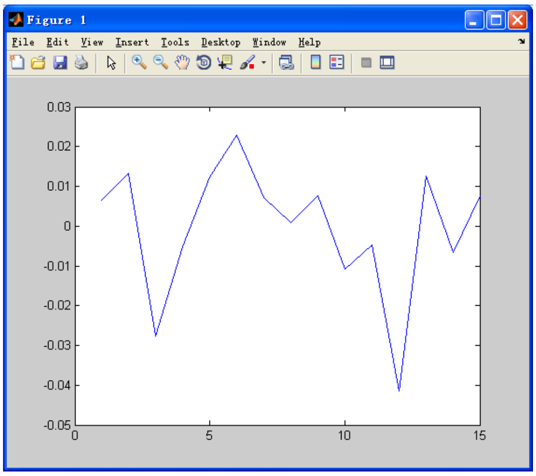

Figure 6. BP Neural Network Deviation Testing Results

Then we process anti-normalization and get the practical fraction deviation. The result is shown in figure 7.

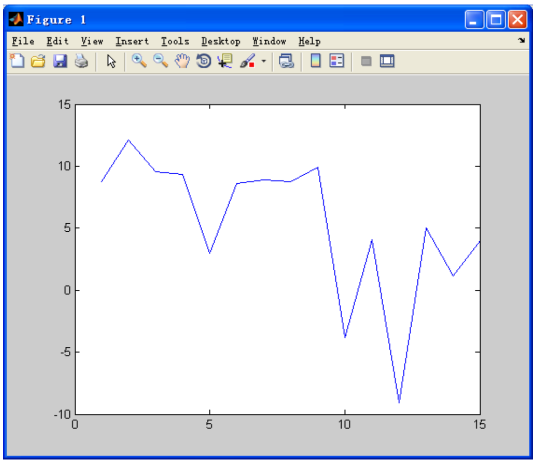

Figure 7. Neural Network Practical Grade Deviation

Then we predict the students should be attributed to the poor performance clusters or not according to all the courses' result. We add a row of data in the original data. Poor performance clusters are set to 1 otherwise set to 0. So this column will be the forecast data. As before, we use one class of students' data to predict and the others to train the network.

The final testing result is shown in figure 8:

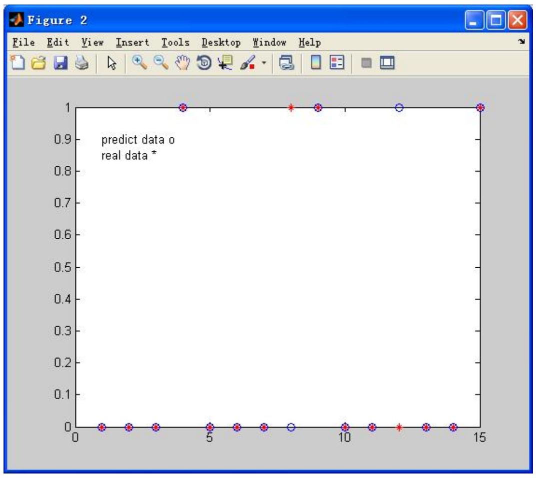

Figure 8. BP Neural Network Early-warning Result

Among them, the hollow circle represents the predicted results and the star represents the actual data. Their overlapping means that prediction is correct otherwise is false.

B. Experimental Analysis of Early- warning

In order to obtain better results, we need to change the parameters constantly when training BP neural network, especially the number of the hidden layer element. We repeated constantly by setting different parameters in the experiment and then we obtain a good result when the final number of the hidden layer element is 9, the maximum training number is 2000. Network's error value meets the qualification after the 27th training. Then the network simulation error is between -0.03 to 0.04 and the range of testing error is -0.04 to 0.02. It can be seen, the results of network training is better. Then we process anti-normalization on students' grade and obtain the practical fraction deviation result. The range of it is -10 to 12. The reason why this range is relatively large, we analysis, is that our early-warning factor is mining by the poor students' grade and our test data contain students of other clusters. This will inevitably affect the final prediction error of students' grade. But we can still warn the student who may not has an ideal grade according to this prediction result. It can remind him may face problems and he needs to review the leading courses. Let him work harder in his future study and prevent him to accessing an undesirable program comprehensive design grade.

By the cluster analysis, we have already known each student's cluster belongs to. So, we can predict whether a student will be clustered to the poor performance cluster or not based on the 13 courses' grade. Then we achieve the Students' Grade early-warning purposes. It can be seen from the final test results, the accuracy is 86.7% with the acceptable range. Then it can send out a warning to the poor performance cluster students. So it can remind them their undesirable grade which may affect their future graduation or employment.

V. CONCLUSIONS

The core section of this paper is the improved data mining algorithms based on an early warning system of college students, which includes detailing historical data area with Clustering algorithm, mining early-warning factor of students' achievement with Association Rules and the Construction of BP neural network prediction model. In K-Means algorithm, the initial cluster center is random distribution. That causes fluctuation effects on cluster result. In response to K-Means algorithm's flaw, combining with the practical application of early warning about students' grade, an optimization algorithm is proposed. The experiment shows that the optimized algorithm makes up for deficiencies of the algorithm and improves the stability of the algorithm as well as ensures the accuracy of the original algorithm. Against the complexity of early warning factor about students' grade and the original algorithm's, this paper puts forward an improved algorithm New_Apriori algorithm and analyzes experiment result. Experiment result shows in the new algorithm, the amount of data access has been reduced significantly and efficiency has been improved. Basing on the early warning factor about students' grade, we build a BP neural model, to mine warning information. Give warning information to some students, who may get poor grade or may be attributed to the class with poor performance. Remind student early and help him to avoid causing more serious consequences.

ACKNOWLEDGEMENTS

This research was supported by China National Key Technology R&D Program (2012BAH20B03), National Nature Science Foundation (31101078),Beijing Nature Science Foundation(4122016, 4112013).

REFERENCES

[1] Vo Thi Ngoc Chau; Nguyen Hua Phung; , "A Knowledge-Driven Educational Decision Support System," Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on , vol., no., pp.1-6, Feb. 27 2012-March 1 201

[2] Hall P. Beck and William D. Davidson. ESTABLISHING AN EARLY WARNING SYSTEM： Predicting Low Grades in College Students from Survey of Academic Orientations Scores[J] . Research in Higher Education．2001,42(6)：709-723

[3] Deirdre Billings． Early Warning Systems：Improving Student Retention And Success[C]．Proceedings of the 15th Annual NACCQ, Hamilton New Zealand July,2002

[4] Zhang Yuntao, Gong Ling.Principles and techniques of data mining[M]．Electronics Industry Press．2004,4：3—40

[5] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M].Mechanical Industry Press.2007,3：1一23

[6] Zhu Ye,Ye Gaoying. Improvement of Apriori Algorithm in Association Rule Mining. Modern Electronic Technique,2008,v31(18).

[7] Xianmin Wei; , "Student Achievement Prediction Based on Artificial Neural Network," Internet Computing & Information Services (ICICIS), 2011 International Conference on , vol., no., pp.485-487, 17-18 Sept. 2011

[8] Margaret H.Dunham. Data Mining Tutorial[M].Tsinghua University Press.2006,3：107-122

[9] Chen Jian. Supporting the Knowledge Management Model of the Distributed Information System- Applied Network Technology and the Decision-making Support System of Data Warehouse.[N].China Information Review.2001(4)：59 一 60

[10] Homayounfar, P.; , "Process mining challenges in hospital information systems," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.1135-1140, 9-12 Sept. 2012

[11] Zhang Tie-lin; Jia Ke-bin; , "Design and Implementation of IDC Engine Room Intelligent Management and Control System Based on Data Mining," Computer Science & Service System (CSSS), 2012 International Conference on , vol., no., pp.1366-1369, 11-13 Aug. 2012

[12] Tan Ran, Lu Zhengqiu, Yan Xinping. Similarity-based distributed data mining model research.Application Research of Computers.Mar.2008, v25(3).

[13] Zhu Ye,Ye Gaoying. Improvement of Apriori Algorithm in Association Rule Mining. Modern Electronic Technique,2008,v31(18).

[14] Li Sha; Gu Qiao-lun; , "The design of R/M integrated supply chain risk early warning system," Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on , vol., no., pp.430-433, 15-17 July 2011

[15] Chen Changxi; Zhang Hongfu; Sa Renna; Wu Hanning; , "Study on early-warning system of pork price based on BP neural networks," Information Science and Engineering (ICISE), 2010 2nd International Conference on , vol., no., pp.4709-4712, 4-6 Dec. 2010

Lijuan Zhou received the BTech degree in Computer Application Technology from the Heilongjiang University in 1991, the MS degree in Computer Application Technology from the Harbin University of Science And Technology in 1998 and the PhD degree in Computer Application Technology from the Harbin Engineering University in 2004.

She is a professor of database system and data mining at the Capital Normal University. She has conducted research in the areas of database systems, data mining, data warehousing, Web mining, object-oriented database systems, and artificial intelligence, with more than 30 journal or conference publications. Her primary research interests are in OLAP, data mining, and data warehouse.

Yuyan Chen received the Bachelor of Engineering degree in Electronic Information Engineering from the Capital Normal University in 2011 and she is currently studying for a master's degree at the Capital Normal

University. Her mentor is Professor Lijuan Zhou and her main research fields are data warehouse and data mining.

Shuang Li received the Bachelor of Engineering degree in Computer Science and Technology from the Hebei Normal University in 2007 and the MS degree from the Capital Normal University in 2011. Her mentor is Professor Lijuan Zhou and her main research field is data mining. He has published three papers in the international conference.