Study and Application of an Improved Clustering Algorithm

Lijuan Zhou
Capital Normal University, Information Engineering College, Beijing, 100048, China
Email: zhoulijuan87@gmail.com

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

***Abstract*— This paper, combined with the characteristics of the early warning about students' grade, represents an optimization algorithm in order to solve the random selection from the initial clustering center of results to cause major influence this volatility defects .It has integrated into the open source WEKA platform. The optimized algorithm not only guarantees the accuracy of the original algorithm, but also improves the stability of the algorithm.**

***Index Terms*— data mining, cluster analysis**

I. INTRODUCTION

As the data volume of database increases constantly, in the process of data mining [1, 2], one data mining time is longer, more rules are mined out. Finally the user will face a mass of rules. Generally, the users are not interested in the potential rules of the overall datum, but some implicit ones. When a general algorithm is mining total data, the mining time increases relatively, so some rules will be hardly found out from the entire rules in which the user is not interested. And probably some rules can't be mined out because of the 'dilution' of the entire data. In this way, the efficiency reduces and useful knowledge can't be got. Therefore before the mining of the potential rules, the data area needs to be thinned according to the user's interests. In the practical application of the students' grade early warning[3,4], using cluster analysis[5] in the data pretreatment stage, firstly cluster the students' grade, secondly thin the data area, thirdly, process the correlation analysis according to the user's interests in specific data. This correlation analysis narrows data set's range dramatically in this process, which makes the mining efficiency improve. In other words, this way combines two mining methods effectively, processes association rules data mining on the basis of clustering. The initial cluster centers of clustering K-Means Algorithm are selected randomly, which causes volatility influence to the clustering results. Aiming at this defect, an improved algorithm of selecting initial cluster centers is put forward. The experimental result shows that the improved algorithm increases its stability in the precondition of guaranteeing the accuracy rate.

II. K-MEANS ALGORITHM

There are four common clustering algorithms: partitioning algorithm, hierarchical algorithms, large database clustering and clustering to classification attribute [6]. 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[7]. It's the most classical 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:

$$m_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij} \quad . \tag{1}$$

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

doi:10.4304/jsw.7.6.1234-1241

research. In view of this, aiming at fixing problem of initial cluster centers, we put forward an optimized k-means algorithm in this paper.

III. SEEK INITIAL CLUSTERING CENTER USING OPTIMIZED ALGORITHM

There are two important concepts in optimization algorithm:

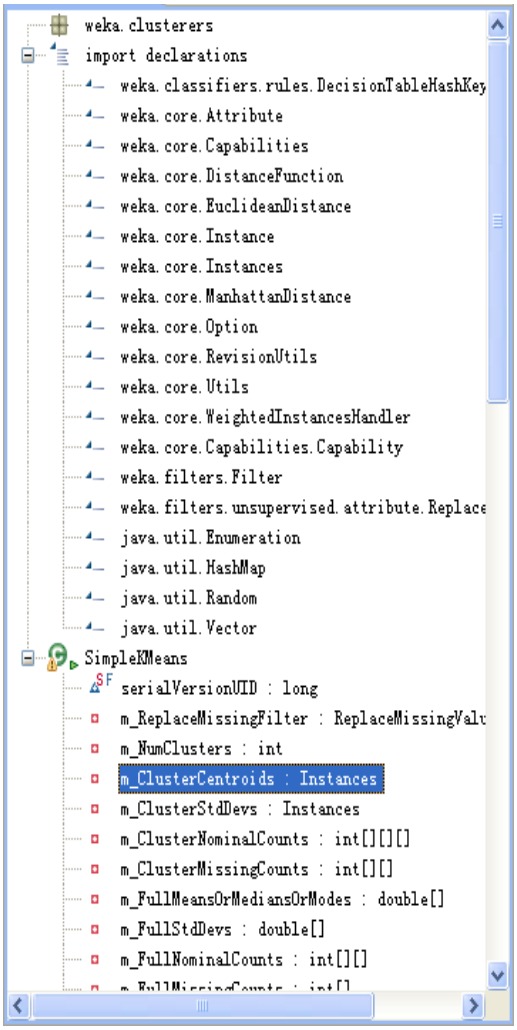

Figure 1. Structure of SimpleKMeans in WEKA

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 figure 1.

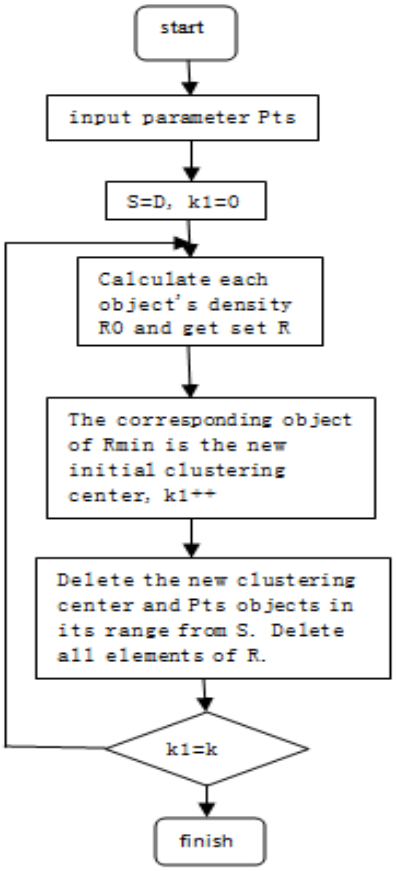

Figure 2. The flow chart of the optimized selecting initial cluster centers algorithm

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. This paper achieved the improved algorithm on WEKA-3.6.0. The code in WEKA is open for us. We can view and analysis source code in package weka cluster after unzipping the weka-scr.jar file in the installation directory into eclipse. This paper improved Simple KMeans on the platform

according to the processes in fugure1. The structure of Simple KMeans on the platform is showed in figure2.

The m_ClusterCentroids marked in figure 2 is a member variable to store cluster centers. It is randomly assigned in SimpleKMeans. Part of the code is as follows:

```
Random RandomO = new Random(getSeed());
int instIndex;
HashMap initC = new HashMap();
DecisionTableHashKey hk = null;
Instances initInstances = null;
if(m_PreserveOrder)
  initInstances = new Instances(instances);
else
  initInstances = instances;
for (int j = initInstances.numInstances() - 1; j >= 0; j--)
{
  instIndex = RandomO.nextInt(j+1);
  hk = new
DecisionTableHashKey(initInstances.instance(instIndex),
initInstances.numAttributes(), true);
  if (!initC.containsKey(hk))
 {

m_ClusterCentroids.add(initInstances.instance(instIndex)
);
        initC.put(hk, null);
  }
  initInstances.swap(j, instIndex);
  if (m_ClusterCentroids.numInstances() ==
m_NumClusters)
{
        break;
  }
   }
```

It is the code of selecting k initial cluster centers. This paper improved this part to emphasize. Improved code of selecting initial cluster centers is as follows:

```
Instances S = new Instances (instances);
Instances initInstances = new Instances(instances);
int k1=0;
double [] R;
int [][] indexofP;//record the object in the range of
every object density parameter-index
int Pts=P;
for(;k1<k;)
{
getDensity(S,Pts,R,indexofP);//obtain density
parameter in the range of Pts of every object in S
   instIndex=getMinR(R,S);//minimum of R
m_ClusterCentroids.add(initInstances.instance(instIndex)
);
k1++;
S.delete(instIndex);//delete the cluster centers object
from S
for(int i=0;i<Pts;i++) //delete Pts objects in the cluster
centers zone from S
        S.delete(indexofP[instIndex][i]);
}
```

The definition of getDensity(Instances S,int Pts,R,indexofP) is as follow:

```
void getDensity(Instances S,int Pts,double [] R, int
[][]indexofP){
      for(int i=0; i < S.numInstances();i++){
            double [] distance;
            int l=0;
            for(int j=i+1; j < S.numInstances();j++)
      {
      double dist =
m_DistanceFunction.distance(S.instance(i),S.instance(j));
      distance[l]=dist;
      indexofP[i][l++]=j;
             }
            for(int m=0; m < l-1; m++){
      int min=m;
      for(int mm=m+1;mm < l; mm++)
      {
      if(distance[min]>distance[mm])
      {
      min=mm;
      double temp = distance[min];

      distance[min]=distance[mm];

      distance[mm]=temp;
      int temp1=indexofP[i][min];

      indexofP[i][min]=indexofP[i][mm];

      indexofP[i][mm]=temp1;
                              }

                        }
            }
      R[i]=distance[Pts-1];
}
}
```

The definition of function getMinR(double [] R,Instances S) is as follow:

```
int getMinR(double [] R,Instances S)
{
       int index=0;
       double min=R[0];
       for(int i=1; i< S.numInstances();i++)
{
            if(R[i]< min) {
                  min =R[i];
                  index = i;
            }
       }
       return i;
 }
```

The purpose of improving the algorithm is to reduce the volatility that randomly selecting the initial cluster centers impact the clustering results. In the source code of

'Random RandomO = new Random (getSeed());', we can see that random number is influenced by parameter 'seed'. This paper experimented with this. The students' grade is clustered into 4 clusters, and we gave each 'seed' different number. The proportion of the 4 clusters in the clustering result changes a lot. Table Ⅰ is the proportion result of different seeds in SimpleKMeans in the WEKA platform.

TABLE I.
PROPORTION RESULT

seed	Cluster 0	Cluster 1	Cluster 2	Cluster 3
5	19%	19%	40%	22%
10	30%	25%	22%	23%
20	11%	39%	5%	45%
30	25%	23%	15%	37%
100	37%	21%	21%	21%
…	……	……	……	……

From the table we can see that the accuracy is higher when seed is 5. Then we experimented with improved Algorithm. The smaller range of constant Pts ensured the stability of the algorithm. At the same time, different Pts choose different initial cluster centers. In this way, we avoided local optimal solution and ensured the accuracy of the algorithm. The students represented by the selecting initial cluster center of improved algorithm are the models of 4 levels students which are classified as "A", "B","C" and "D" by their grades. This describes that the improved algorithm not only reduces the volatility of the old algorithm but also ensured the stability of the result.

Ⅳ. EXPERIMENT RESEARCH

A. *Data preparation*

Data sources from a 4-year-grade of the students in a normal class. After cleaning and converting the data sources, we get the data of cluster analysis as table Ⅱ shows:

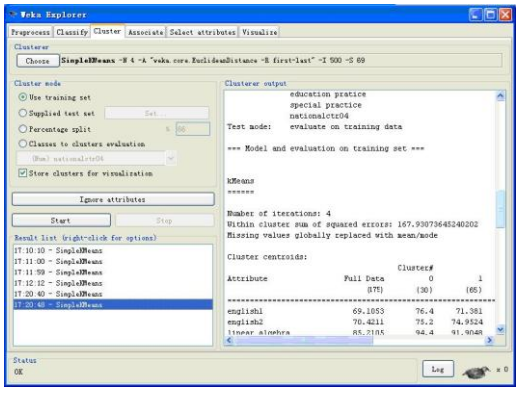

Figure 3. Clustering results

TABLE II.
STUDENT SCORE DATABASE OF CLUSTER ANALYSIS

id	english 1	english2	linear algebra	C	…	military training	…
1	60	86	52	22	…	2	…
2	82	73	99	93	…	3	…
3	73	73	98	88	…	1	…
4	73	69	88	82	…	2	…
5	64	70	72	66	…	2	…
6	68	57	81	67	…	2	…
7	70	68	89	81	…	2	…
8	65	64	86	84	…	2	…
9	71	64	83	77	…	2	…
10	69	67	100	71	…	2	…
…	…	…	…	…	…	…	…

B. *Improved algorithm analysis with 4 clusters*

By clustering analysis, we get the clustering result as figure 3:

TABLE III.
NUMBER AND PROPORTION OF EXAMPLES OF 4 CLUSTERING CLUSTERS

Cluster	Instances	Percentage
cluster0	30	17%
cluster1	65	37%
cluster2	40	23%
cluster3	40	23%

We do the statistical analysis with the 4 cluster centers in each course and observe the distribution situation of maximum and minimum values of 4 cluster centers.

TABLE IV.
DISTRUBUTION SITUATION OF MAXIMUM AND MINIMUM VALUES OF 4

Cluster	**maximum**	**minimum**
cluster0	20	1
cluster1	21	0
cluster2	0	39
cluster3	1	2

It's easy to see that the grades of cluster0 and cluster1 are better, cluster2's is worse and cluster3's is general in table Ⅲ and table Ⅳ.

Then we use a graph to observe the grade's distribution of the 4 clusters in each course. The cluster result is in the

last of properties. Click it to view the statistics and histogram. The histograms from the left to the right stand for cluster0 cluster1, cluster2, cluster3.As figure 4 shows:

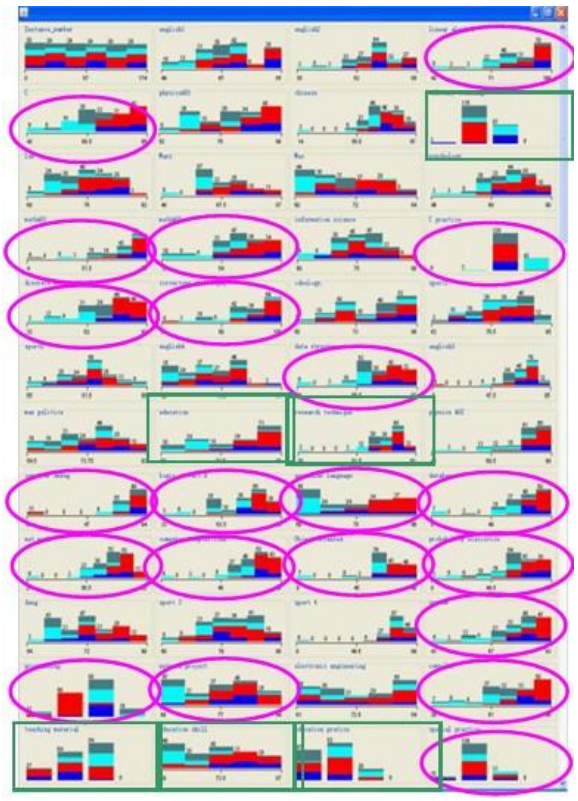

Figure 4. Distribution of the course "College English level 1"

Select "english1" to analyze the distribution of 4 clusters in the course "College English level 1", as figure 5 shows:

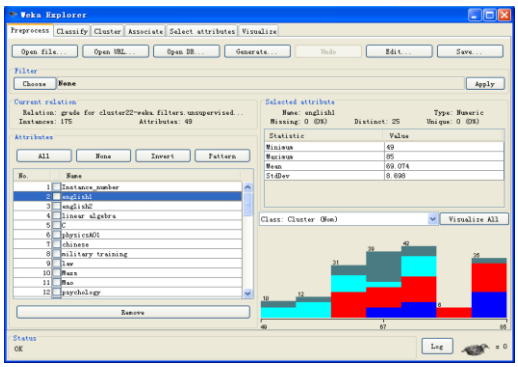

Figure 5. Distribution of the course "College English level 1"

We can see from the histogram that the score of 'College English Level 1' ranges from 49 to 85. The score on the right is higher than in the left. The overall result is not high because there are so many students in the middle. The light blue that stands for the cluster2 occupies a large proportion of low scores and a small proportion of high scores. Both of the red that stands for the cluster0 and the blue that stands for the cluster1 occupy a large proportion of high scores. The gray that stands for the cluster3 occupies different proportions in each part. The largest proportion is in the middle and small proportion is in the other parts. This proves that the cluster center of cluster3 is in the middle.

Next, we analyzed the situation of computer professional courses. First of all, we choose 'C programming language' which is the earliest course they learnt. We showed the analysis of C programming language. As figure 6 shows:

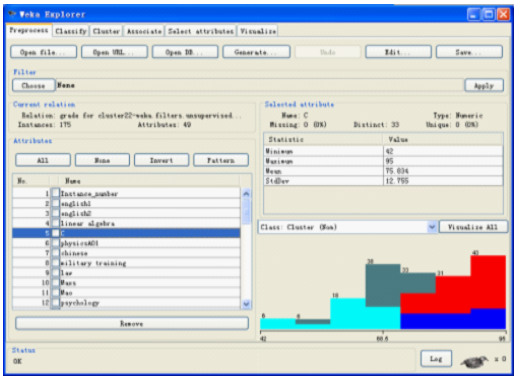

Figure 6. Distribution of the course "College English level 1"

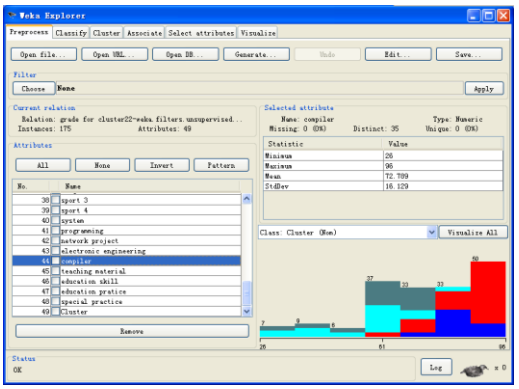

Figure 7. Distribution of the course "Compiler Principle"

The histogram shows that the scores of all the students range from 42 points to 95 points. The light blue that stands for cluster2 almost occupies all the low scores. The red that stands for cluster0 occupies a large proportion of high scores. The blue that stands for cluster1 occupies many high score students too. And the grey that stands for cluster3 is still a large proportion of middle scores. Therefore, we suspect that the scores of cluster0 are better than that of cluster1.

Then we choose the representative professional courses, 'Compiler Principle'. We showed the analysis of Compiler Principle. As figure 7 shows:

The histogram shows that the scores of all the students range from 26 points to 96 points. And most of the scores are relatively high because high-score students occupy the largest proportion of all the students. The light blue that stands for cluster2 still occupies a large proportion of low scores and a small proportion of high scores. The red that stands for cluster0 almost occupies all the high scores. The blue that stands for cluster1 and the gray that stands for cluster3 are the same situation as they are in figure6.

Therefore, the result confirms our guess that the scores of cluster0 are better than that of cluster1.

These histograms in figure 8 are the situation of all the courses:

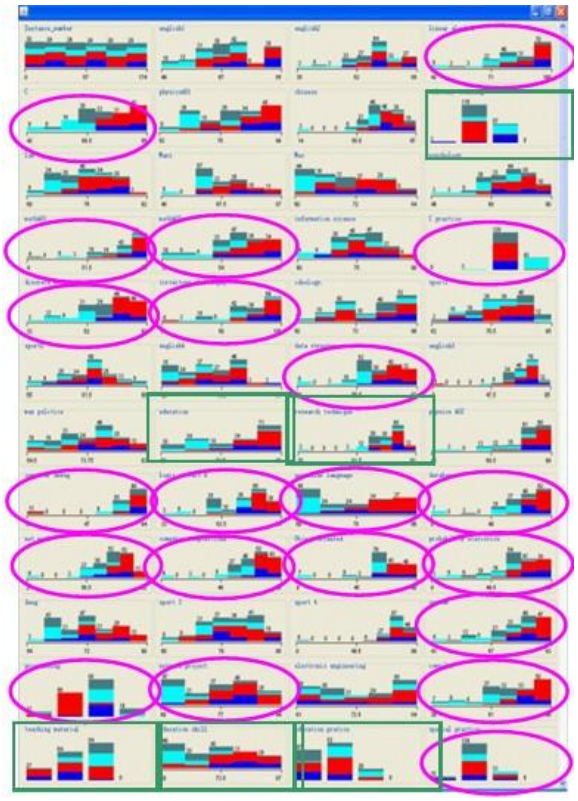

Figure 8. Distribution of the course "College English level 1"

The courses that their histograms marked by rectangles are the computer profession classes while the courses that their histograms marked by ellipses are pedagogy curriculums. They both affect students' achievement. By analyzing a lot of histograms of different courses, we confirmed our earlier conjecture further: the scores of cluster0 are higher than cluster1. And the whole situation is that the score of the light blue that stands for cluster2 is lowest, the score of the red that stands for cluster0 and the blue that stands for cluster 1 are higher, and the score of the gray that stands for cluster3 is in the middle.

We checked the result with the fact of students, and it verified our conjecture.

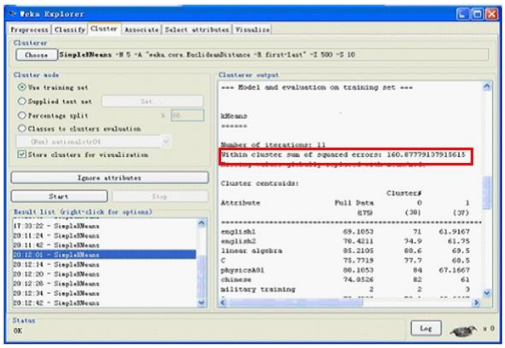

Figure 9. The result of 5 clustering clusters

C. *Improved algorithm analysis with 5 clusters*

Some teachers are prefer to classified students into 5 levels such as "A", "B","C" , " D" and "E" by their grades. So we try to set the number of expected clustering cluster to 5, and let's see whether it can produce a more accurate result from figure 9. The histograms from the left to the right stand for cluster0 cluster1, cluster2, cluster3 and cluster4. As figure 10 shows. And the table 5 shows the proportion of the instances in different cluster.

TABLE V. NUMBER AND PROPORTION OF EXAMPLES

OF 5 CLUSTERING CLUSTER

Cluster	Instances	Percentage
cluster0	30	17%
cluster1	37	21%
cluster2	43	25%
cluster3	40	23%

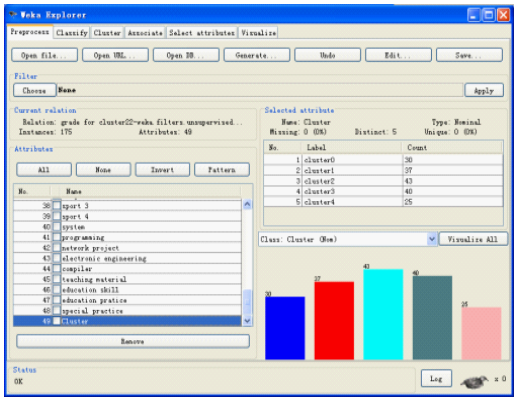

Figure 10. 5 clusters histogram clustering results

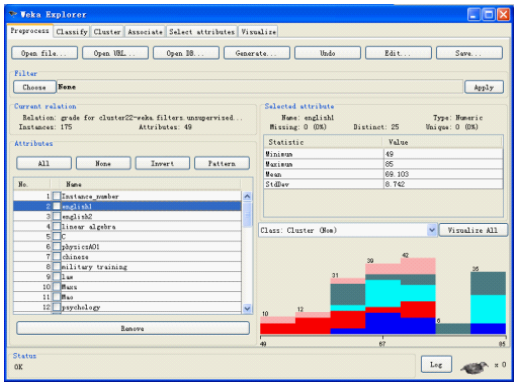

Figure 11. Distribution of the course "College English level 1" of 5 clusters

Figure 11, 12 and 13 are the distributions of different courses when the number of cluster is 5. From these figures we can see that: the red that stands for cluster1 is worse and occupies a part of low scores, its

Characteristics is not obvious. The focus of research of students' achievement early warning is low scores data zone, so we need the data has distinguished feature and be representative. On the other hand, the grey that stands for cluster3 and the light blue that stands for cluster2 are both represent the high block, the blue that stands for cluster0 and the pink that stands for cluster4 are both in the middle, and we can't distinguish which one is higher.

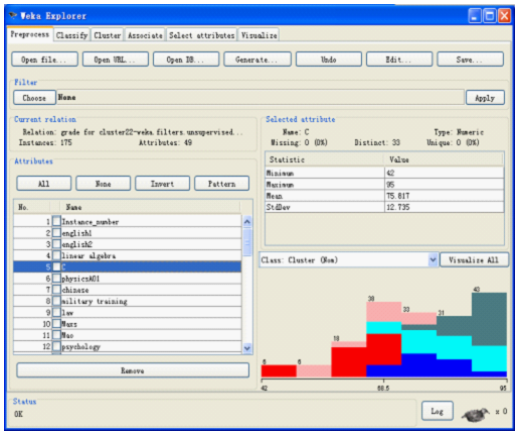

Figure 12. Distribution of the course "College English

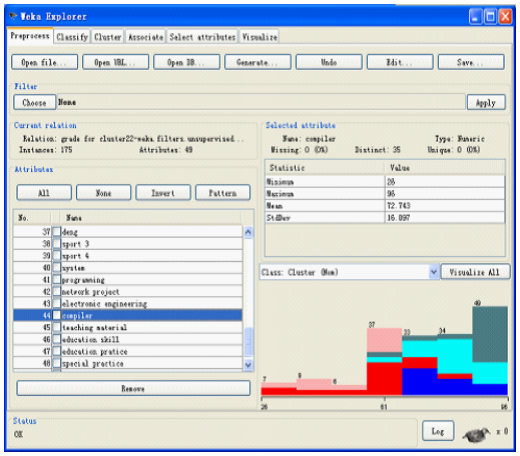

Figure 13. Distribution of the course "Compiler Principle"

According to the students' actual situation, we choose the result of the number of cluster is 4. The students that cluster 2 stands for in this result are the data area we choose for students' achievement early warning .

V. CONCLUSION

This paper is mainly about the use of clustering algorithm in data mining to pretreat students' achievement. At first, we introduce the concept of the k-means algorithm, and then we propose an optimized algorithm to make up for deficiencies of the algorithm and integrate into the open-source WEKA platform. Then we clean and convert the data sources and we cluster analysis by using the improved algorithm through the weka platform. 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.

ACKNOWLEDGMENT

This research was supported by China National Key Technology R&D Program (2009BADA9B00),

This research was supported by National Nature Science Foundation (61070050) ,

This research was supported by Beijing Educational Committee science and technology development plan project (KM20111002818),

This research was supported by the Open Project Program of Key Laboratory of Digital Agricultural Early-warning Technology, Ministry of Agriculture, Beijing, 100037.

REFERENCES

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

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

[3] 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

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

[5] Yang Xia-ling, Nie Yong-hong. Application of the clustering analysis in the forecast of employment of graduates [J].Journal of Guang Xi University of Technology.2005,16(4):82-84

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

[7] 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)：59－60

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 Capital 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.