Advances in Intelligent Systems Research, volume 133
2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

The Research of Grade Prediction Model Based on Improved K-means Algorithm

Yongguang Zhang, Hua Wang* and Hongyang Li
Information Engineering College of Capital Normal University, Beijing, China
*Corresponding author

***Abstract*—Grades reflect how well you learnt in courses. This paper introduce a model to predict student grade-data with a refined K-means clustering algorithm. K-means clustering algorithm based on the normal distribution is proposed to overcome the flaws that caused by using Euclidean distance algorithm to measure the similarity between objects. Experiment results show that K-means clustering algorithm based on the normal distribution is more accurate than classical K-means clustering algorithm in grade-data prediction.**

Keywords-k-means algorithm; grade prediction; similarity measurement

I. INTRODUCTION

Clustering is a kind of partitioning which based on the special feature of datasets. Clustering can make objects which are in high similarity into the one of classes and make obvious gaps between different classes. Clustering is a method to partition the unsigned objects into different classes which means there is no any rule or standard to partition those objects before clustering start. So clustering belongs to unsupervised learning. After clustering, there is an evaluation about the accuracy of the result. There are varieties of algorithm to reckon the distance between objects. What calls for special attention is that which algorithm is more suitable to cluster the given data. Whether we selected a suitable algorithm to reckon the distance between objects directly influence the accuracy of the clustering result.

II. THE CLASSICAL K-MEANS CLUSTERING ALGORITHM

In 1976, Mac Queen proposed K-means algorithm to our view. K-means is a classical algorithm which is widely used in data mining. Compatibility for most of data type is the key advantage of it. According to the given variable-k, K-means algorithm partition all given objects to k clusters.

First, the k-means method choose k objects as the initial clustering centers from the data set. Then, reckon the distance with the similarity measuring function between every objects and every centers. Objects are classified based on the distance-rule. The rule is that the object belongs to the cluster as long as the distance between the object and the clustering center is the most minimum distance.

Next step, calculate the mean of every objects within a cluster. Select the mean data within clusters as the new clustering centers and continue to partition the data by calculate the distance between objects and new clustering centers until the newest clustering centers are equal to the prior clustering centers. The equality also means that the criterion function is convergent (K-means use error sum of squares as the criterion function). Figure 1 is the algorithm flow table.

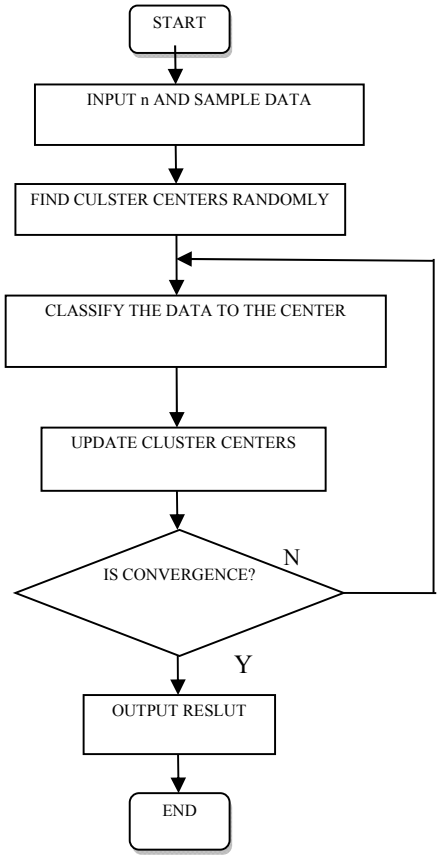

FIGURE I. K-MEANS ALGORITHM FLOW TABLE

III. THE REFINED K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm measure the similarity rely on calculating the distance between objects. There are a lot of distance measuring algorithm which perform well in clustering low-dimension data. However, there are few algorithm that is suitable to deal with high-dimension data which means using distance measuring algorithm to partition high-dimension objects is tend to result in the inaccuracy. In grade analyzing work and grade prediction work, k-means algorithm which use traditional distance measuring algorithm to calculate distance is not suitable to cluster the grade data which is a kind of high-

dimension data (because each students learn numerous courses in school so they get numerous grades).

In classical K-means clustering algorithm, the result is tend to be influenced more by the course which grade variance is higher. Sometimes, the difference of grades cannot reflect how much level-difference between students because features of some courses make student grades hardly to be very low or extremely high. So, sometimes only focus on score of the course is not enough to judge that how well the student learned in the course.

Rank, ignoring the variance and mean value of the course, can reflect the level of a student. But the real rank information cannot be found in the grade list so the paper quote a formula to estimate the rank.

Based on the normal distribution, the paper put forward a refined k-means algorithm which can overcome the prior problem that only consider the score. The refined algorithm calculate distance between objects and clustering centers with normal distribution algorithm instead of using Euclidean algorithm to calculate the difference between grades.

The distance calculated by normal distribution algorithm reflect the rank of the student grade. Ignoring the difference in score between each student, we can calculate the rank of student course (the rank-evaluation formula is showed in the following paragraph). With the rank-estimate formula, we then can get the normal distribution value which is the evidence for calculating the distance between objects.

The following formulas show the mathematical theory of the K-means algorithm based on normal distribution.

$$F(A,B) = |1/n\sum_{i=1}^{n} f_{ai}(\mu, \sigma) - 1/n\sum_{i=1}^{n} f_{bi}(\mu, \sigma)|$$

$$fai(\mu, \sigma) = \frac{1}{\sqrt[2]{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

Value μ is the mean value of scores of the ith course. Value σ is the standard deviation of the ith course. Value X is the score of the ith course of Student a.

Fai(μ , σ) means the rank position of the "ith" score in the normal distribution(rank-evaluation).

F(A, B) is the similarity between students. The lower F(A, B) is, the more similar between students are.

First, calculate fai. It means that map every score of the ith courses to estimate the normal distribution value (This step covert the 0-100 range score to the 0-1 range data). For instance, Student A get 83 score in Higher Mathematics. We calculate the mean value μ and the variance σ from his classmate's Higher Mathematics scores. X is 83. We put them into the fai formula then get the value between 0 and 1 which reflect the performance of student A in Higher Mathematics.

Then, calculate F(A, B) with the prior formula in chapter 3. F(A,B) is the distance between objects. The new distance value F(A, B) reflects the similarity between students. Compare to calculate the Euclidian distance value directly, the new distance value reflecting the similarity is more accurate because the new distance algorithm consider the rank of the score into calculation process not the score itself only.

Finally, we can start to cluster. The following paragraphs show the clustering process.

- Step A) Input value k and data D. Value k is the number of clustering centers. Choose k objects from D as the initial clustering center randomly.
- Step B) Calculate the distance (based on normal distribution algorithm) between objects and each clustering centers. Classify each object to the nearest (the lowest distance value) clustering center.
- Step C) Update the value of clustering centers. Calculate the mean value of objects in each center then replace the prior clustering center to the new clustering center.
- Step D) Return to B until the values of prior clustering centers are equal to the values of new clustering centers. Equations

IV. PERFORMANCE PREDICTION MODEL BASED ON THE IMPROVED K-MEANS CLUSTERING ALGORITHM

The process of grades prediction: Cluster the sophomore student first-year-course grade with K-means clustering algorithm based on normal distribution to train the model. Input the new student grade data, calculate the distance (with new algorithm) between the data and each prior clustering centers. Choose the nearest clustering center as the key-center. The dataset samples are the data which are classified into the same cluster together with the key-center. Those samples are used for prediction. Choose sample's second-year-course grade. Then calculate the mean value of the second-year-course grade. The mean value is the result of the performance prediction to the grade-input student. The Figure III shows the process with a flow chart.

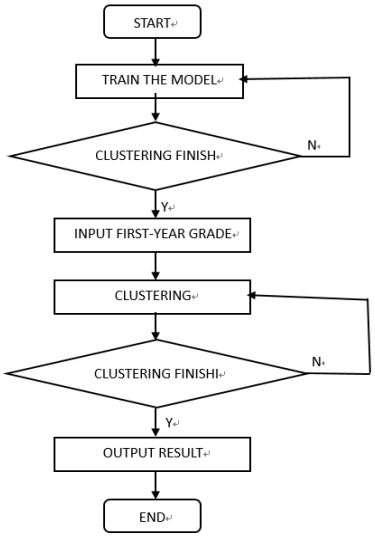

FIGURE II. REFINED ALGORITHM PREDICTION FLOW CHART

V. THE TEST AND THE ANALASIS OF THE TEST RESULT

The test use 3 batches of student course grade (initial batch) as the initial data. Those student course grade is used for clustering as the sample for following prediction work. The first-year-grade is selected to be the data of clustering. The table 1 shows a part of data sample.

TABLE I. PART OF DATA IN THE INITIAL BATCH

	SUBJECT 1	SUBJECT 2	SUBJECT 3	SUBJECT 4
1101000001	99	91	79	70
1101000002	78	87	80	66
1101000003	77	66	74	67
1101000004	87	87	80	68
1101000005	72	72	68	72
1101000006	73	73	74	67
1101000007	87	87	72	74
1101000008	85	85	77	69
1101000009	90	90	81	67

Another batch of student course grade (test batch) is used for testing (being predicted). This batch of data is also used for calculating the accuracy of the test result (prediction work). Table II shows a part of first-year-grade of the test batch data.

TABLE II. PART OF FIRST-YEAR GRADE DATA IN THE TEST BATCH

	SUBJECT1	SUBJECT2	SUBJECT3	SUBJECT4
1141000001	70	81	76	71
1141000002	92	95	74	71
1141000003	90	93	92	68
1141000004	89	92	95	73
1141000005	84	72	89	69
1141000006	90	93	81	73
1141000007	94	89	93	70
1141000008	70	77	72	62
1141000009	77	60	77	67

Firstly, cluster those 3 batches of data (initial batch). Use K-means algorithm based on normal distribution to classify the first-year grade data of 23 courses. Then, calculate the distance between the students in the test batch and the clustering centers in the initial batch. After distance calculation, classify them into the nearest cluster. Calculate the mean value of 23 course grade in each cluster. Finally, as the prediction of test batch, the result is the second-year grade mean value of the nearest cluster in the initial batch.

To evaluate the accuracy of the prediction, we calculate the difference between the prediction result and the real grade (the second-year grade) of the test batch. The Table III shows a fraction of the real grade (the second-year grade) of the test batch and the Table IV shows a fraction of the prediction result.

TABLE III. PART OF SECOND-YEAR GRADE DATA IN THE TEST BATCH

	SUBJECT 5	SUBJECT 6	SUBJECT 7	SUBJECT 8
1141000001	79	76	77	90
1141000002	98	94	90	75
1141000003	93	81	86	90
1141000004	74	90	88	87
1141000005	73	79	75	91
1141000006	98	99	91	94
1141000007	86	81	96	82
1141000008	84	72	84	81
1141000009	70	50	60	74

TABLE IV. PREDICTION RESULT OF TEST BATCH

	SUBJECT5	SUBJECT6	SUBJECT7	SUBJECT8
1141000001	79.8	77.5	76.3	93.0
1141000002	96.5	95.5	91.5	73.9
1141000003	92.2	82.3	88.2	88.3
1141000004	75.9	87.2	91.0	89.5
1141000005	75.5	80.6	76.8	96.8
1141000006	97.1	98.5	89.6	90.9
1141000007	85.3	81.6	95.5	84.0
1141000008	85.6	73.3	83.1	81.8
1141000009	72.2	51.2	60.2	77.9

If the difference value between the real grade and the corresponding result is less than 3, we define that the result is eligible. If any grade predictions of the student are eligible, we define that the prediction for this student is accurate. The table 5 shows the accuracy rate and other details of the prediction work.

TABLE V. THE ACCURACY RATE OF THE PREDICTION

SAMPLE SIZE	ACCURATE	INACCURATE	RATE
84	67	17	79.76%

The paper tend to propose a model that can make grade prediction for students according to their past grade. The prediction can drew the attention of students to pay more time on learning the coming low grade course. As the table 5 shows that 84 students was predicted and there are 67 of them are predicted accurately. Furthermore, there are 17 students are predicted as inaccuracy. In summary, the accuracy rate of the prediction model is 79.76%.

Nowadays, the BP neural network and the Bayesian network are two main method to predict the student grades.

In training process, BP neutral network include input layer, hidden layer and output layer. It is a multilayer forward learning method. The method can find the image of output data and input data by one-to-one map. In the text, we use the

first-year grade and second-year grade as the training samples to train the BP neutral network prediction model. Then, use a new batch of first-grade data as the test data to get the prediction result.

Bayesian grade prediction model is built by the algorithm of edge-removing. It also predict the second-year grades according to the first-year grade data of students.

The following summary paragraphs are mainly about the analysis of comparison between classical BP neutral network and Bayesian network and K-means clustering algorithm based on normal distribution.

Summary 1:

Predict the five grades of College Physics, Advanced Mathematics, Probability and Mathematical Statistics, Data Structures and Algorithm and Computer Networks (These five courses cannot be selected at the first year.) form students' first-year grades with 3 different kinds of algorithm. Compare the prediction result with the real grade the student got in these five courses. As we defined before, if the error between the prediction result and the real grade is less than 3, we see the prediction result as an eligible result. The Figure 3 shows that the eligibility rate of the different 3 kinds of prediction algorithm result.

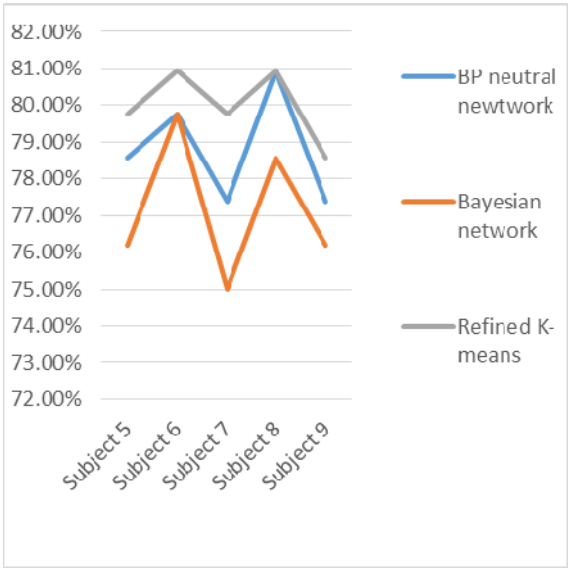

FIGURE III. ELIGIBILITY RATE OF DIFFERENT KINDS OF ALGORITHM

As the Figure III showed, the prediction results of K-means clustering algorithm based on the normal distribution are closer to the real grade compare to the other algorithms.

Summary 2:

Predict all course grades then calculate the accuracy rate. If all the prediction results of a student are eligible, we defined that the prediction for this student is accurate. The accuracy rate of 3 kinds of prediction algorithm is showed in the Figure V.

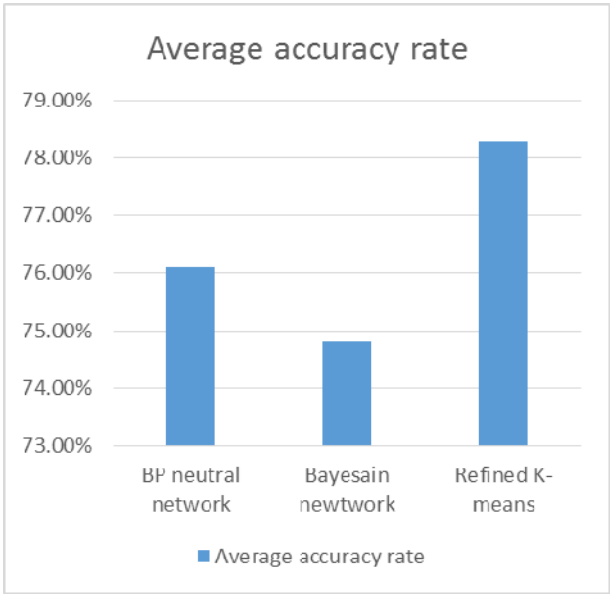

FIGURE IV. AVERAGE ACCURACY RATE OF DIFFERENT KINDS OF ALGORITHM

As the Figure IV showed, this is the mean accuracy rate of the prediction result with 3 different kinds of algorithm. The X axe present the kind of algorithm. The Y axe is the number of the mean accuracy rate. The mean accuracy rate of refined K-means prediction algorithm reaches 79.76% which is the highest rate of these 3 kinds of algorithm.

VI. CONCLUSION

The paper tend to propose a grade prediction model with K-means clustering algorithm based on normal distribution. The result of experiment shows that the refined k-means is the better algorithm for prediction compared with BP neutral network and Bayesian network. The prediction result provide an evidence for students' study in the future. However, it just a theoretical prediction which means the final real grade rely on how much efforts the student will make

REFERENCES

[1] Steinley, Douglas, and Michael J. Brusco. 2011. "Choosing the Number of Clusters in K-Means Clustering." Psychological Methods 16, no. 3: 285-297.

[2] Kuo, R. J., N. J. Chiang, and Z.-Y. Chen. 2014. "Integration of Artificial Immune System and K-Means Algorithm for Customer Clustering." Applied Artificial Intelligence 28, no. 6: 577-596.

[3] Hu, Meng-han, Qing-li Dong, Bao-lin Liu, and Pradeep K. Malakar. 2014. "The Potential of Double K-Means Clustering for Banana Image Segmentation." Journal Of Food Process Engineering 37, no. 1: 10-18. Business Source Complete, EBSCOhost (accessed July 24, 2016).

[4] Hartigan, J.A., Clustering algorithms. 1975.

[5] Jiang, Feng, Guozhu Liu, Junwei Du, and Yuefei Sui. 2016. "Initialization of K-modes clustering using outlier detection techniques." Information Sciences 332, 167-183.

[6] J. van der Geer, J.A.J. Hanraads, R.A. Lupton, The art of writing a scientific article, J. Sci. Commun. 163 (2000) 51-59.

[7] Kuo, R. J., N. J. Chiang, and Z.-Y. Chen. 2014. "Integration of Artificial Immune System and K-Means Algorithm for Customer Clustering." Applied Artificial Intelligence 28, no. 6: 577-596.