Network teaching system based on a clustering analysis algorithm

Lanzhong Wang

Shandong University
Shandong, People's Republic of China

ABSTRACT: To provide learners of network-based teaching programmes with targeted tutoring and to obtain better learning outcomes, data mining technology was applied to network education. Students in an on-line programming language course from the Network Education School at Shandong University in China were subjects in this study. An analysis and evaluation of the students' academic performance was conducted using an improved clustering algorithm. The results show that the algorithm has good clustering performance and computational speed, and it may be a useful evaluation tool in distance education systems.

INTRODUCTION

Network teaching is an important learning method based on the use of modern network technology. Teachers are an educational resource separated from the students and they adopt the role of learning promoter [1]. Positioning the student's learning and providing targeted tutoring has become a problem that needs an urgent solution [2].

This research used data mining technology to analyse network education, and an evaluation of students' academic performance was conducted by using the improved K-mean clustering algorithm on educational statistics. As a result, teaching staff can adjust their teaching according to a student's learning characteristics.

CLUSTERING ANALYSIS

Definition of Clustering Analysis

Clustering analysis refers to an analysis that divides a physical or abstract object set into groups of a similar nature. The goal is to classify data based on this similarity [3]. In a given collection of elements D, each element has n observable properties. D is divided into k sub-collections by some algorithm.

The dissimilarity degree of elements in each sub-collection is required to be as low as possible, while the dissimilarity degree of elements between different sub-collections is required to be as high as possible [4]. The essence is a global optimum problem. The distance between vectors A_i and A_j (A_i= {A_{i1}, A_{i2}, ..,A_{in}}, A_j = {A_{j1}, A_{j2},.., A_{jn}}) in n-dimensional space Sn is given by:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^2} \tag{1}$$

K-means Clustering Algorithm

K-means clustering divides data based on a minimum error function [5]. K points are first selected as the centre of the K clusters. The remaining elements are used to calculate the dissimilarity degree of the K cluster centres according to Formula (2). The results are divided into clusters with similar dissimilarity degree. Then, the arithmetic mean of each dimension of all elements is selected from the cluster, and the central values of the K clusters are recalculated. The process is iterated until all the elements are re-clustered according to the new centres. Figure 1 shows the control flow for the K-means algorithm.

$$M_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij} \tag{2}$$

The performance of the K-means clustering algorithm greatly depends on the initial clustering centres. The shortcoming of selecting initial clustering centres randomly is that the classification obtained from these initial clustering centres severely deviates from the global optimum classification. When the clustering value is greater, such shortcoming is more obvious. Further, the time complexity is high, while the elasticity is poor in the process.

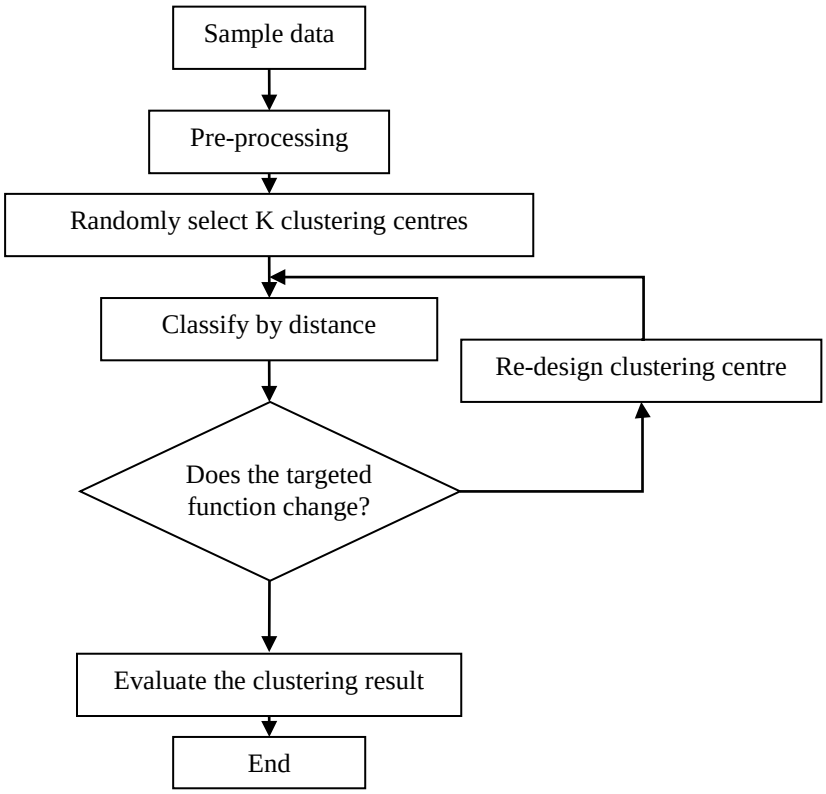

Figure 1: Flow chart for the K-means clustering algorithm.

IMPROVING RESEARCH WITH K-MEANS CLUSTERING ANALYSIS ALGORITHM

To speed-up the clustering algorithm, the density threshold-based grid clustering method was added to the K-means clustering algorithm. The grid of sample space is divided applying the speed of grid clustering; and the first clustering is completed by smoothly filtering out noise. For discrete data with a small density threshold, K-means clustering algorithm is applied to implement the second clustering and so on until the condition is satisfied.

Grid-based Clustering K-means Algorithm

For the bounded domain set P = {X_1, X_2, … X_n}, in the n dimensional space S = X_1*X_2*…*X_n, the input to the algorithm is an n dimensional point set Q = {q_1, q_2, …q_n}, where q_i = {q_{i1}, q_{i2}…,q_{in}}, q_{ij} refers to the j dimensional component of the i^{th} point. The density D of each grid cell is set as the number of points of the unit. In terms of setting of density threshold, literature suggests improvement on the traditional DBSCAN (density-based spatial clustering of applications with noise) algorithm [6]. The density value D(C_i) of N points with the highest density in the grid cell is selected. The calculation is:

$$Minpts = \frac{\left[\sum_{i=1}^{N} D(C)^2\right]^{\frac{1}{2}}}{N} \tag{3}$$

Under normal conditions, clustering is based on D(C_i) in descending order. If the difference value between D(C_{i+1}) and D(C_i) is greater, it will consider a transition to have occurred. At this point, set N = i. The specific steps of the algorithm are as follows:

1. Divide each dimension of the n-dimensional space into r non-related and equal size intervals to form Rn grid cells. The length calculation of each grid cell in the i dimension is $\delta_i = (h_i - l_i)/p$ and, then, the *j* interval section:

$$Iij = \left[l_i + (j-1)\delta_i, l_i + j\delta_i \right] \quad (4)$$

2. Map points with data concentrations onto the unit set; calculate the density $D(C_i)$ of each grid cell.

3. Classify grid cells according to the density threshold: the density, which is greater than the high density cell of Minpts will be marked directly; the density, which is smaller than the low density cell of Minpts is treated as isolated discrete data awaiting the next processing.

4. Repeatedly select grid cells, and its neighbouring cells, until all density cells are clustered. According to Formula (2), calculate the value Gi (0) of K clustering centres, and take it as the initial clustering centres.

5. As to discrete data with low density cells, calculate the distance to the clustering centre dies (arc), when it obtains the minimum value, a$\in$ C_i, repeat this until all the discrete data clustering are done.

6. Recalculate second clustering gravity Gig(1), for example, if $|Gi(1)-Gi(0)| < \varepsilon$, then, the clustering is completed or re-circulate through the K-means clustering method until $|Gi(m)-Gi(m+1)| < \varepsilon$ is satisfied.

COMPARISON OF CLUSTERING ALGORITHMS

An experiment was performed, which adopted the classical dataset Iris of the UCI machine learning repository (a collection of databases, domain theories and data generators). There were 150 samples. Each sample had four properties, divided into three categories.

The GKC (grid-based and K-means clustering), DBSCAN, and traditional K-means clustering algorithms, were tested. The time for the GKC algorithm is mainly to position the data density area and calculate the initial clustering centre. The time complexity can be divided into order O(2d*r) and O(K*I*M), where I and M are the iteration times and discrete data quantity.

Table 1 shows the comparison of clustering times for the three kinds of algorithm. It can be seen that the GKC algorithm is better than the other two algorithms in terms of iteration speed.

Table 1: Contrast of clustering times.

Experiment times	GKC	K-means	DBSCAN
1	0.498	0.523	0.505
2	0.496	0.571	0.497
3	0.493	0.604	0.511
4	0.491	0.566	0.529
5	0.488	0.603	0.490

Clustering performance uses a purity degree for measurement. The purity degree of a cluster E_{ij} is equal to the intersection of cluster i and classification j [7]. As shown in Figure 2, GKC is better than the K-means or DBSCAN algorithm in terms of purity, and the purity degree fluctuates less, meaning it has better stability.

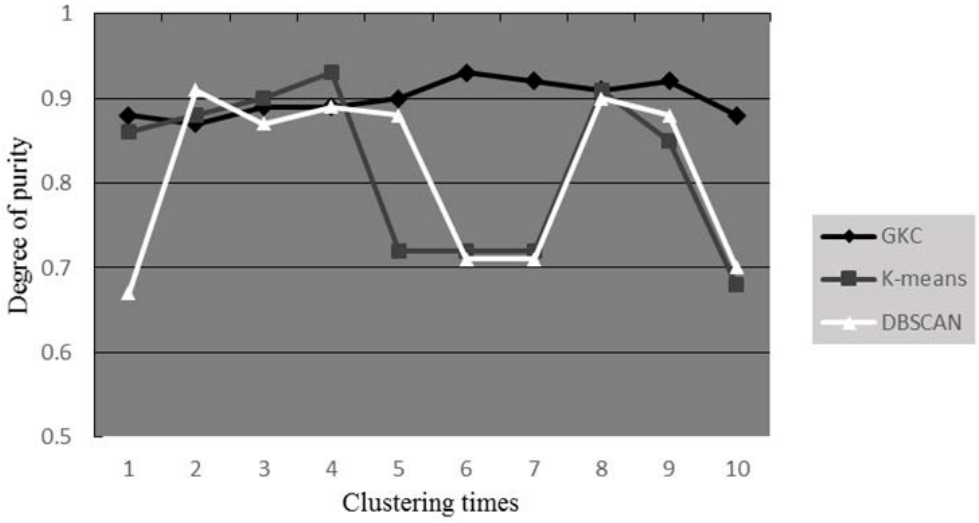

Figure 2: Purity for the three algorithms.

APPLYING CLUSTERING ANALYSIS TO DISTANCE TEACHING

The work reported here used an on-line course of the Network Education School of Shandong University, entitled C programming language, to illustrate how the clustering algorithm applies to, and plays its role in, the evaluation of learning. The number of students engaged in the on-line evaluation of the course totalled 86. The knowledge points are numbered as indicated in Table 2. Clustering analysis was conducted on the students' scores using the GKC clustering algorithm. The specific process was:

1. Input the students' test scores and initialise data.
2. Clustering analysis using GKC to work out clusters and clustering results.
3. Analysis of the clusters.

Step 3 - analysis of the clusters - included average scores of cluster members, the number of cluster members, knowledge points where students were weak, and knowledge points where students were strong.

Table 3 shows the clustering results from the clustering analysis.

Table 2: Knowledge points for the C programming language course.

Knowledge point no.	1	2	3	4
Knowledge point	Input and output	Function	Pointer	One-dimensional array
Knowledge point no.	5	6	7	8
Knowledge point	Two-dimensional array	Array and pointer	Character string	Variable scope
Knowledge point no.	9	10	11	12
Knowledge point	Macro Replacement	Structure, community	Bit operation	Documents

Table 3: Clustering results.

Cluster no.	Average score	Knowledge points - students weak	Knowledge points - students strong	Number of the cluster members
1	54	4,5,6	9	15
2	81	9,10	1	12
3	65	1,7	8	31
4	70	10,12	2	28

These results provide learning suggestion(s) and guidance. The students could carry out targeted learning based on the analysis results; thus, improving weak areas.

Figure 3 describes the basic structure of a network-based teaching system, and the role of clustering analysis. So far, the analytical function of the clustering algorithm has not been embedded into the distance education system, which will be realised subsequently by future researches.

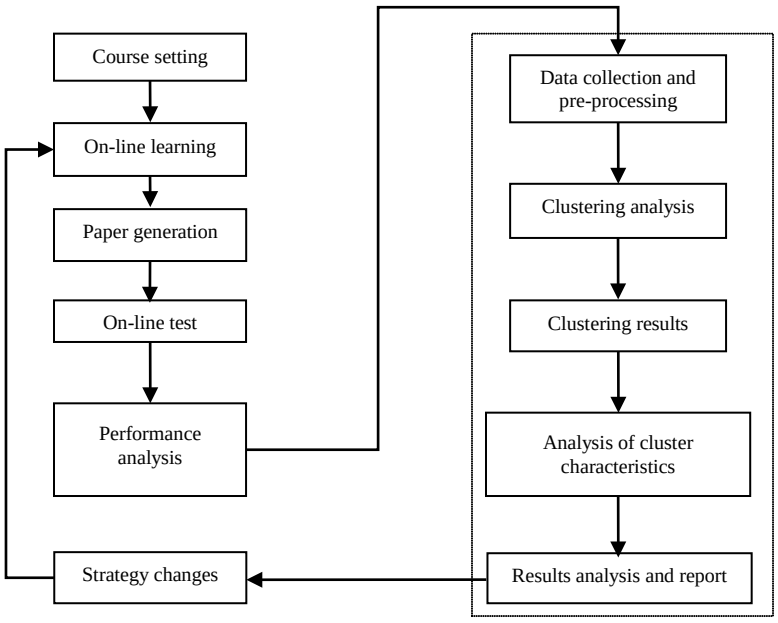

Figure 3: Clustering analysis in a network-based teaching system.

CONCLUSIONS

To better utilise the teaching resource of network-based teaching to achieve the required teaching objectives, this work involved in-depth research on the function of clustering algorithms and their role in learning assessment.

A specific example of the use of a clustering algorithm in network-based teaching was considered. Improvements have been made to the K-means clustering algorithm by applying grid clustering. This overcomes the uncertainty caused by K value randomness, as well as the cluster loss caused by the traditional grid clustering method.

The application of such technology will be conducive to improving teaching in distance education systems.

REFERENCE

1. Chen, M., SWOT analysis and strategies to support college physical education through network teaching. *World Trans. on Engng. and Technol. Educ.*, 12, **4**, 671-674 (2014).
2. Liu, Z. and Wu, Y., Design of a university aerobics teaching network information platform (ATNIP). *World Trans. on Engng. and Technol. Educ.*, 13, **1**, 34-37 (2015).
3. Ping, D., Application of clustering analysis in teaching evaluation. *J. of Hunan Institute of Engng.: Natural Science Edition*, 01, 35-38 (2010).
4. Dongming, T., Research on clustering analysis and its application. *University of Electronic Science and Technol.*, 08, 22-25 (2010).
5. Di, T., Xizhi, Z. and Xiaohang, L., A clustering algorithm based on the extension of K-means. *J. of Henan Institute of Educ.: Natural Science Edition*, 16, 2, 26-28(2007).
6. Yanfeng, X., and Wei, M., Research on an improved ant colony clustering algorithm based on cloud model. *Automation and Instrumentation*, 33, **11**, 79-81 (2010).
7. Zhang, X., Research on the application of clustering analysis algorithms in network teaching system. *Bulletin of Science and Technol.*, 29, **4**, 106-108 (2013).