The Clustering Analysis Method of the Learning Characteristics Based on the Virtual Learning Community

Yan Cheng, Jian Hua Xie, and Zhi Ming Yang

***Abstract*—With the rapid development of social economy and the Internet, the network education is becoming a way of teaching which has a wide application range and covering larger area. Virtual learning community (VLC) is a combination of computer technology, psychology, pedagogy and other multi-disciplinary research field and actually a new model of network education. However, the teaching data of VLC are often disorderly, fragmentary, mixed and its value is also not easy to detect. The using of data mining technology will solve this kind of problems and bring many unexpected benefits support the teaching of the VLC. This paper reports on the analysis of learning behavior of the VLC and how to extract the feature vector of learning. The fuzzy c-means clustering algorithm is applied to analyze the learning behavior and divide the students of the VLC by the feature of them. Then some targeted teaching guidance can be made for each group. This kind of grouping strategy is to be found feasible and achieved good effect by simulation experiment.**

***Index Terms*—Fuzzy c-means clustering algorithm，learning characteristics, virtual learning community(VLC).**

I. Introduction

With the popularity of computer and network vigorously, network education is becoming an important way for people to learn. The development of network education offers the diversity that people can choose way he love to study on the Internet, also let people feel the convenience of network learning. As a new development direction of network education, the attention has been gradually drew and the expect to the VLC is higher and higher.

There are so many researches on the VLC about how to construct the VLC and the relationship between the learner, teacher and the community. From the perspective of the social network analysis method, L. Wang [1] did some research on the relationship between the participants in the VLC. A theoretical foundation has been made for the how to forecast and design a better the function of the virtual learning community. F. J. Ma [2] designed a technology model of virtual learning community under the web 2.0 environment. From the aspects of the integrity of system, the dynamics of intelligence, the learning mode and the knowledge management, Y. C. Gan and Z. T. Zhu [3] studied the learning architecture of the knowledge building and the development of swarm intelligence. X. Y. Wen [4] advanced a way to build the VLC by the use of social semantic web. S. H. Gong [5] think that the construction of a learning community should not paying so much more attention to the hardware as the software is also need to be considered. At the same time, that the formation of the members to the community interaction and the community culture is also very important .To build the effectiveness evaluation index system ,she consider the technology and platform interface, the resources and target orientation, the cultural exchanges and interaction, the emotion and situation atmosphere .Viktor Freiman studied the mathematics and other subjects in the VLC through the new method, and opened up a new way of interaction[6]. Ben Daniel used the hybrid method to find out if the interactive content of students reflects the basic elements of the community [7].

After reading a large number of papers, the research about the construction of the VLC and the evaluation index system from the macroscopic angle had obtained a certain achievement. Against this background, in this paper, we develop a logic-based representation that how to analyze the characteristics of learning in the learning process and how to group the learner by the fuzzy c-means clustering algorithm through the micro level.

Then some practical teaching strategies and learning tasks will be made by related experts after analyzing the learning characteristics of the learners in the process of study. And we think this will be a good way for the learners to know themselves and to be better through the plans which the related experts made.

II. The Extraction of the Study Feature Vector

VLC is not only a network platform, but also a learning organization. It supports many kind of teaching service for the learners who have various learning background. In the virtual learning community, there are all sorts of study resource. And the background of learner, also their learning behavior, is different. That how to provide good teaching services for all kinds of learners is a focus in the study of virtual learning community. In this paper, in order to provide a reasonable teaching strategy for learners, we analyzes the learners' learning characteristics and discuss how to make a division for the study group .Taking the students of ideological political education classes of grade 2014 in Jiangxi Normal University as the research object, therefore, we designed a model of teaching strategy based on learning characteristic vector (see Fig. 1)

Manuscript received September 5, 2015; revised October 21, 2015. This work was supported by National Natural Science Foundation of China (NSFC, Grant No. 61262080), Jiangxi Province Science and Technology Support Major Project (20151BBE50121).

Yan Cheng is with Tongji University. She is also with Jiangxi Normal University, Nanchang, Jiangxi, China (e-mail: chyan88888@jxnu.edu.cn).

Jian Hua Xie and Zhi Ming Yang are with Jiangxi Normal University, Nanchang, Jiangxi, China (e-mail: 971383331@qq.com, 709830862@qq.com).

doi: 10.18178/ijiet.2017.7.1.843

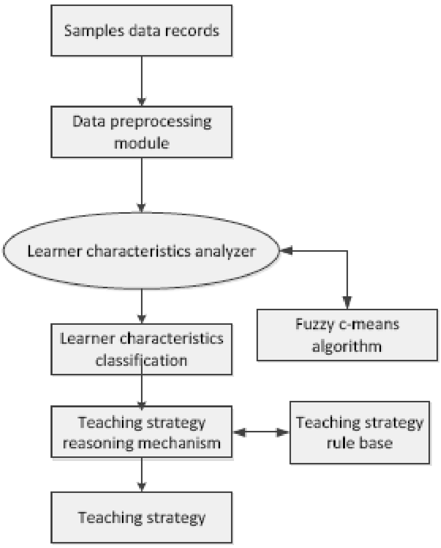
Samples data records
Data preprocessing module
Learner characteristics analyzer
Fuzzy c-means algorithm
Learner characteristics classification
Teaching strategy reasoning mechanism
Teaching strategy rule base
Teaching strategy

Fig. 1. A model of teaching strategy based on learning characteristic vector.

A. Data Preprocessing

1) Data cleaning

The main task of data cleaning is to supplement the missing value, smooth noise data, identify or remove outliers to solve the inconsistency of data. Virtual learning community platform store a lot of information of learning what is multidimensional data. In the process of processing of these data, due to a variety of reasons, such as machine or man-made, some data may deviate from the real value and will have more influence on the analysis of the results. We treat this kind of data as outliers and will delete them.

2) Data integration

The goal of the data integration is to gather the data from multiple data sources in different parts of the unified. Due to the different data source, the expression of the same data or the properties of the same data is different. We should integrate the data to avoid the inconsistency and redundancy.

3) Data conversion

Data conversion, in other word, is to simplify the data. Such as learners' grades and final grades are the reflections of learners’ characteristics of knowledge level. We can handle the data using a one-dimensional knowledge level comprehensive show the grades and final grades.

B. Learning Characteristic Analysis Module

Both traditional education and Internet education, the levels of knowledge and independent learning are the typical indicators of learners' characteristics. In addition, according to the characteristics of the network education, learners' collaborative learning enthusiasm will serve as another important study characteristic. Therefore, we chose the learner's knowledge level, independent learning and collaborative learning enthusiasm to analyze. Then we analyze its impact factors and their weights to the levels of knowledge, independent learning and cooperative learning enthusiasm according to the actual situation of Virtual learning community. Among them, the learners' learning characteristic vectors are defined as follows.

Definition 1:

Suppose there are *n* learners, each the dimensions of the learners' learning characteristics is m. Then the characteristic vector of learners is $S_i = (S_{i1}, S_{i2}, \ldots, S_{im})$. S_i represents the i^{th} learners’ learning characteristic vector, S_{im} represents the m^{th} learning characteristic vector to the i^{th} learner (*i* is no greater than *n*).

TABLE I: LEARNERS' CHARACTERISTICS AND WEIGHT

Learning characteristic	Knowledge levels S_{i1}	Independent learning S_{i2}	Cooperative learning enthusiasm S_{i3}
Usual scores (P1)	Weight Q1		
Final scores (P2)	Weight Q2		
Practice times in the homework module (T1)		Weight Q3	
The learning times in the course module (t2)		Weight Q4	
The test times in examination module (T3)		Weight Q5	
The learning times in the resource sharing module (X1)			Weight Q6
The times of uploading data (X2)			Weight Q7

TABLE II: PART OF THE LEARNER'S LEARNING CHARACTERISTICS

Student ID / Learning characteristic	Knowledge levels (points)	Cooperative learning enthusiasm（times）	Independent learning （times ）
14*****001	83.91	1.7	42.1
14*****054	40.00	2.2	12.8
14*****003	83.33	2.9	7.7
14*****004	95.65	1.0	23.1
14*****005	88.51	9.9	67.9
14*****006	90.90	1.9	8.2
14*****007	53.89	1.4	31.1
14*****008	44.44	7.2	14.7
14*****009	59.88	4.0	9.1
14*****010	66.29	0.7	2.8
14*****011	73.97	1.6	8.8
14*****012	90.38	1.3	16.4
14*****013	84.93	0.7	15.9
14*****014	73.88	1.7	25.0
14*****015	90.07	1.4	38.2
14*****016	67.27	0.7	22.6
14*****018	73.98	0.7	19.6
14*****019	69.44	0.7	4.2
14*****020	87.96	2.0	43.8
14*****021	81.57	3.1	28.4
14*****022	62.81	1.9	20.3
14*****023	59.23	13.4	35.0
14*****024	93.69	1.6	30.2
......	......	......	......

In this paper, we do not discuss the weight evaluation, and the weight value determined by assuming [8]. The learner's knowledge level depends on the scores they get at the usual time and final. And we assume each weight is assumed as

*Q*1=0.3, *Q*2=0.7;The independent learning consists of three parts ,the practice times in the homework module , the learning times in the course module and the test times in examination module . And each of their weight is assumed as *Q*3=0.3, *Q*4=0.4, *Q*5=0.3; Collaborative learning enthusiasm consists of the learning times in the resource sharing module and the times of uploading data ,and each of their weight is assumed as *Q*6=0.3, *Q*7=0.7. In this paper, the learning characteristic vector is $\mathrm{S_i} = (\mathrm{S_{i1}}, \mathrm{S_{i2}}, \mathrm{S_{i3}})$, $\mathrm{S_{i1}}$ represents the knowledge levels, $\mathrm{S_{i2}}$ represents the independent learning, $\mathrm{S_{i3}}$ represents the cooperative learning enthusiasm.

Table I shows the weight of each learning characteristic. Based on the autonomous learning website of Jiangxi Normal University computer course. And this learning website is study community based on Moodle (Modular Object-Oriented Dynamic Learning Environment) which has recorded the action of the students in this website .Through the Moodle we can get the data we need in the Table I.

Here we take the students of ideological political education classes of grade 2014 in Jiangxi Normal University as the research object. And there are 60 students in this class. We will consider one course named fundamentals of computer as an example that 5 invalid records and 55 valid records. According to the setting of the weight to each influence factor, the learners learning characteristics had been calculated and shown as below (see Table II).

III. The Fuzzy C-Means Algorithm Grouping

Fuzzy C-means algorithm (FCM) is a kind of common fuzzy classification method which Bezdek had used in pattern recognition [9]. According to the similarity of the learning characteristics, we grouped the students by fuzzy c-means algorithm and calculated the center characteristic vector of each group. That’s the main work this paper do.

After the learners have been divided into groups, the domain experts make some corresponding teaching strategies for targeted teaching so as to realize the community learners' overall improvement.

A. The Description of Learner Clustering Analysis Problem

After classifying learners by learning characteristics, each type of learners have a clustering center, respectively form a vector of clustering center. Each type center vector of the learners' learning characteristics defined as follows.

Definition 2:

The learners will be classified by their learning characteristics through cluster algorithm. And the cluster center vector of each group is defined as $Z_k = (Z_{k1}, Z_{k2}, ..., Z_{km})$, Z_k represents learners' overall learning characteristics of group *k*, Z_{km} represents the value of the *m*th learning vector in the group *k*. The value of *k* is equal to the number of category *c*.

Clustering analysis was carried out on the learners, namely, the learners are divided into category *c*. The learner's each classification corresponds to a *c* × *n* membership degree matrix: $U = \{u_{ij}\}$, u_{ij} represents for the membership degree between learners with cluster subset S_j . $V = (v_1, v_2, ..., v_c)$ represents the clustering center set for all learners.

B. The Objective Function of Learners Clustering Analysis

The fuzzy c-means algorithm [10] based on clustering analysis was carried out on the learners' learning characteristics. The objective function of clustering algorithm is as below.

$$Min\ J_{\mathrm{m}} = (U,\ V) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d_{ij}^2 \qquad (1)$$

To make formula (1) minimum, namely, the sum of square of the distance between each learner sample to the center of the cluster. In the formula 1, $d_{ij} = \|x_j - v_i\|$ represents the distance between the *j*th learner S_j and the *i*th clustering center v_i , m represents the weighted index, usually value to 2 [11]. The weighted index controls the distribution of the membership degree and the degree of fuzzy clustering. The larger the value is, the higher the fuzzy degree is. u_{ij} represents the membership degree of the learner S_j to cluster center. At the same time, u_{ij} must satisfy the following conditions:

1) $u_{ij} \in [0,1]$, each student for each category of membership degree is between 0 and 1.

2) $\forall i, \sum_{i=1}^{c} u_{ij} = 1$, every learner to the sum of all the c category of membership degree is 1.

3) $\forall j, 0 < \sum_{j}^{n} u_{ij} < n$, for a particular category, there is always at least one learners should be affiliated with it.

C. Learners Membership Degree and the Calculation of Clustering Center

The calculation formula about the learners for each category of membership degree is shown as equation 2. d_{ij} represents the distance between the *j*th learner and *i*th clustering center, $\sum_{k=1}^{c} d_{kj}$ represents the sum of distance between the *j*th learner and *c*th clustering center. Actually *m* value of 2.

$$U_{\mathrm{ij}} = \left[\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}\right]^{-1} \qquad (2)$$

The calculation formula of the learner's clustering center is shown as shown formula 3

$$V_{\mathrm{i}} = \sum_{j=1}^{n}(u_{ij})^{m} x_j / \sum_{j=1}^{n}(u_{ij})^{m} \qquad (2)$$

D. The Steps of Fuzzy C-means Algorithm

Fuzzy c-means algorithm steps are as follows:

Step 1: Determine the number of learner classification *c*, $2 \le c \le n$, *n* represents the number of learners . Given the initial cluster centers $V(0)$ randomly, set the termination value ε of the iteration process. Set the initial numerical iteration times p = 0.

Step 2: According to the formula (2), calculating the value u of the membership degree, then we can get $U^{(p)}$.

Step 3: through to the formula (3), we can calculate the new matrix clustering center $V^{(p+1)}$.

Step 4: We can get J_m through the formula (1). If $\left|J_m^{(p)} - J_m^{(p-1)}\right| > \varepsilon$, then make *p*=*p*+1 and turn to step 2, otherwise stop to calculation.

IV. The Experimental Process

Here with MATLAB2012b simulation process to the 55students is as follows:

A. The Parameters of the Fuzzy C-Means Algorithm Initial Setup

1) The selection of weighted index *m*: Weighted parameter m has a great influence and effect on the performance of the algorithm. The selection of value m should not only ensure the weighted error sum of squares in the class to small, and to ensure the kind of spacing to big [12]. In this experiment *m* value is 2.
2) Set the maximum number of iterations p: the number of iterations is set to 150
3) The terminating condition: set to 1e-6, which is the smallest variable of the weighted error sum of squares in the class to be less than 1e-6, and then stop the iteration.
4) Clustering of group c: Clustering algorithm is an unsupervised learning method [13]. The number of category c in this fuzzy c-means algorithm should be selected according to particular case for particular analysis.

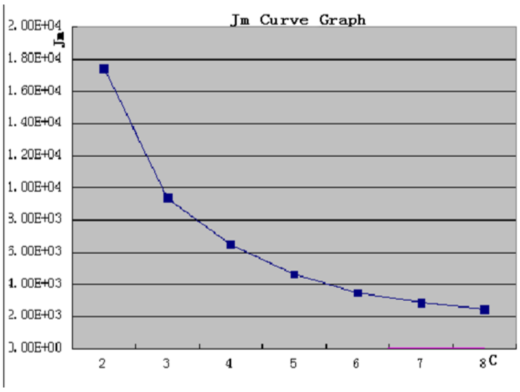

Fig. 2. The relationship between J_m and the category c.

We chose an appropriate number of category after we compare the relationship between category c of clustering analysis with objective function of J_m. The relation between learners category *c* and J_m are shown in Fig. 2 below .The experimental results shows that with the increase of *c*, learning characteristics the weighted error sum of squares (J_m) of the learning characteristic became smaller. When *c* equals 6, it's a turning point, then change is more gently, thus category number *c* is set 6.

B. The Implementation Process of the Fuzzy C-means Clustering Grouping

1) Initialize clustering center learning characteristics. The initial clustering center was selected by center by the system functions selected randomly. Then the appropriate category c can be obtained by the above process.
2) The actual number of iterations. The experiment of the maximum number of iterations is set to 150 times. Actually when the learning characteristics of the recent change of weighted error sum of squares is less than the set termination conditions of 0.00001, the iteration is stopped when the iterate number is 51.
3) Iterative 51 times, the final clustering center of learning characteristics and the membership are determined.
4) Learners are divided into 6 classes according to their membership.

Each learner have a membership degree to the 6 clustering centers, the learner will be divided into one of maximum value membership category.

After the termination of algorithm, learners are divided into 6 categories according to learning characteristic by the fuzzy c-means algorithm, shown in Table III.

TABLE III: The Result of the Group Clustering

Learning characteristic Center feature vector	Independent learning （times）	Cooperative learning enthusiasm （times）	Knowledge levels (points)
Z_1	72.8604	3.8411	81.1484
Z_2	10.7434	2.0922	19.4637
Z_3	14.3021	3.9764	51.2058
Z_4	43.0204	2.3565	82.8676
Z_5	13.5856	2.7441	86.4123
Z_6	20.3907	2.6183	72.6090

The calculation result shows that the 55 students were divided into 6 groups according to the learner's learning characteristic. Then a relatively large group is split into small groups with similar characteristics that targeted teaching strange can be made for them. In this paper, through the detailed teaching strategies what made by the domain experts for each group, learners in this class have a better learning enthusiasm and proved to have achieved a good results.

V. Conclusion

Against the background of network education, this paper analyzes the existing problems in the network education under the new age. By analyzing the characteristics of learning, namely treating the clustering center feature vector as the research object, then we can simply the problem of learning characteristic by constructing the vector of learning characteristic .Then the related domain experts of vector formulate specific learning tasks and goals for each group according to the learning characteristics of each group which have been classified. The next step we will do some research about how to establish a forum for each group and formulate related discussion topics to provide relevant learning resources and so on. In conclusion, we hope the study can provide some related theory and practice support for the further development of network education.

References

[1] L. Wang, "Social network analysis of virtual learning community," *China Educational Technology*, pp. 5-11, Feb. 2009.

[2] F. J. Ma, "Design for the model of the virtual learning community based on web 2.0," M.S. dissertation, Shandong Normal University, 2008.

[3] Y. C. Gan and Z. T. Zhu, "Virtual learning community knowledge construction and the development of the collective wisdom study framework," *China Educational Technology*, pp. 27-32, May 2006.

[4] X. Y. Wen, L. H. Kong, and Z. M. Jiao, "Research on the construction of virtual learning community on the basis of social semantic web," *Modern Education Technology*, pp. 97-101, Oct. 2013.

[5] S. H. Gong and X. F. Cao, "The construction of evaluation index system of virtual learning community," *Contemporary Educational Science*, pp. 15-18+26, May 2015.

[6] V. Freiman and N. Lirette-Pitre, "Building a virtual learning community of problem solvers: Example of CASMI community," *ZDM*, vol. 41, no. 1-2, pp. 245-256, 2009.

[7] B. Daniel and R. A. Schwier, "Analysis of students' engagement and activities in a virtual learning community: A social network methodology," *International Journal of Virtual Communities and Social Networking (IJVCSN)*, vol. 24, 2010.

[8] S. TL, "A scaling method for priorities in Hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 281-342, 1997.

[9] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, NY: Plenum Press, 1981.

[10] X. X. Sun, X. X. Liu, and Q. R. Xie, "The implementation of the fuzzy C-means clustering algorithm," *Computer Applications and Software*, pp. 48-50, Mar. 2008.

[11] X. B. Gao, J. Li, and W. X. Xie, "Optimal choice of weighting exponent in a fuzz C-means clustering algorithm," *Pattern Recognition and Artificial Intelligence*, vol. 13, no. 1, pp. 7-11, Jan. 2000.

[12] G. Y. Gong, Y. C. Mao, X. B. Gao, and S. Y. Liu, "Fuzzy c-mean clustering method for analyzing microarraygene expression data," *Journal of Xidian University*, pp. 291-295, Feb. 2004.

[13] W. Z. Zhao, H. F. Ma, Z. Q. Li, and Z. Z. Shi, "Efficiently active learning for semi-supervised document clustering," *Journal of Software*, vol. 23, no. 6, pp. 1486-1499, Jun. 2012.

Yan Cheng was in February, 1976 in Wu Yuan City. She got the bachelor's degree in 1998, the master's degree in 2005, and the Ph.D. degree in 2010 from Tongji University in Shanghai and she is now a postdoctor of Tongji University in 2015, majored in computer science, engaged in e-learning, virtual learning community and educational technology innovation. From July 2013 to March 2014, she went to the California University as a visiting scholar. Cheng's main research interests include intelligent computer aided education, educational data mining, virtual learning community and e-learning.

Currently she is a postdoctor of Tongji University and a professor in Jiangxi Normal University in Nanchang, Jiangxi province. As a researcher, currently she is the leader of 2 national NSFC (Natural Science Foundation of China) Plan Project and more than 5 provincial projects on IT and education innovations. She is the author of more than 30 scientific papers, won Guanghua Ph.D. scholarship, 3 talks in the global intelligent automation conference, 1 national doctoral academic best paper award. She published 1 academic monograph (Beijing, Science Press, 2014) on educational data mining, 5 computer professional teaching materials.

Dr. Cheng is the senior member of the Systems Engineering Association of China, the review expert of Computer Application Study, one member of Chinese Intelligent Automation, the consulting evaluation expert of small and medium-sized enterprise in Jiangxi province. Dr. Cheng is also the direction leader of the ministry of education "software engineering" discipline and the expert in education informationization.

Jian Hua Xie received the bachelor's degree in June 2010 from Jiujiang University, Jiangxi province. Currently he got the master's degree in June 2015 under the direction of Dr. Cheng. His research interests include complex system modeling, Virtual learning community.

Zhi Ming Yang received the bachelor's degree in June 2013 from Hubei University of Arts And Science, Hubei. Currently he is a master of computer science and technology of Jiangxi Normal University. His research interests include personalized study, virtual learning community and non-cooperative game.