Application of DBSCAN clustering algorithm in evaluating students' learning status

Hui Du
Northwest Normal University
Lanzhou, China
duhuiywy@nwnu.edu.cn

Shiyin Chen
Northwest Normal University
Lanzhou, China
13507021302@163.com

Haopeng Niu
Northwest Normal University
Lanzhou, China
784525493@qq.com

Yongbiao Li
Northwest Normal University
Lanzhou, China
yongbiaoli@yeah.net

***Abstract*—Poor teaching effect is a common problem in the traditional teaching process. This article uses DBSCAN clustering algorithm to classify students and develops corresponding teaching strategies according to different types of students. First, extracting the characteristics of the students as data, including career orientation, self-control index, and knowledge level. Then, using the DBSCAN algorithm to cluster the students. It has been verified that the use of DBSCAN clustering algorithm to grasp the learning status of students has certain guiding significance for teaching activities.**

Keywords-Student* Characteristics*; Student Learning Status; DBSCAN Clustering Algorithm

I. INTRODUCTION

In teaching activities, there are common problems that affect the teaching effect, such as uneven students' level, lack of interest in learning, and easy distraction. Vygotsky believes that the recent development zone refers to the gap between the actual development level and the potential development level of children. The former is determined by the child's ability to solve problems independently, while the latter refers to the child's ability to solve problems under the guidance of adults or working with more capable peers [1]. Interest is the most important and lasting motivation for learning and exploring new things, and it is the prerequisite for acquiring knowledge and ability [2]. The famous Hyde Middle School in the United States also emphasizes "character first" in its school-running philosophy. They believe that if teachers can cultivate students' integrity, self-confidence, pragmatism, curiosity, caring for others, cooperative spirit and other characteristics, students' academic success will naturally come [3]. Students' self-control ability, personality and professional interest are important factors that affect the teaching effect.

There are few researches on the application of clustering algorithms in the education field, and most of the researches focus on K-means clustering algorithm or the data dimension is relatively single. Xiao Lizhong proposed that the K-means clustering algorithm based on particle swarm optimization is more suitable for analyzing the state of students and the K-means algorithm based on genetic algorithm than the K-means clustering algorithm based on the two characteristic dimensions of students' classroom situation and after-school life as the data set [4]. Abualhaj B pointed out that the main disadvantage of K-means is that the number of clusters must be defined in advance [5]. Liusheng Lu proposed an analysis method of student learning status based on density local outlier algorithm. Use student performance as the data source to find out suspicious outliers; the algorithm focuses on finding abnormal students [6].

We choose the DBSCAN algorithm to analyze the learning status of students. The DBSCAN clustering algorithm is different from other classic clustering algorithms. According to the situation in the teaching process, there are the following three points. First, the DBSCAN clustering algorithm is not easy to fall into a local optimal solution. Second, the DBSCAN clustering algorithm does not need to pre-determine the number of clusters. Third, the DBSCAN clustering algorithm can effectively separate noise points. In teaching practice, there are individual poor students, outstanding students and special students, which require special attention from teachers. These students are the "noise points" in the algorithm. We can find these students and pay attention to them through the DBSCAN algorithm. According to the above three points, it is not difficult to see that the application of DBSCAN clustering algorithm in teaching practice has a high degree of adaptability.

With the development and popularization of computer technology in our country, it has greatly promoted the development of related fields, and its application in the field of education has become an inevitable trend of social development [7]. How to apply the rapid development of information technology to students' teaching practice will become a breakthrough in education and teaching.

2021 17th International Conference on Computational Intelligence and Security (CIS) | 978-1-6654-9489-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CIS54983.2021.00084

978-1-6654-9489-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CIS54983.2021.00084

II. STUDENT CHARACTERISTICS DIMENSION

A. Career Orientation

Students are not very interested in learning. It is difficult for students to maintain long-term interest in learning, become bored in the learning process, and refuse to actively participate in the classroom. If there is no continuous and real interest in learning, follow-up learning will become a new problem.

Students' vocational interest is also an important factor in the teaching process. Personality and occupation are very closely related. The matching degree of personality type and occupation type determines the success of the career [8]. Holland believes that people prefer to work with people who have similar professional interests and personalities and prefer to choose their career fields. If a person has a certain personality type, it is easy to be interested in a certain type of occupation, and thus is also suitable for this type of occupation [9]. The difference in personality of each person determines their attitude towards work and the difference in the work they are good at. Based on the development concept that students are human beings, cultivating students according to their professional interests in the teaching process is not only conducive to mobilizing students' interest, but also conducive to the future development of students.

The Holland Career Interest Questionnaire used by the research is compiled with reference to the revised Holland Career Tendency Test Scale [10]. Holland's career interest theory has been widely used in career guidance, social and business fields, and its influence is expanding day by day [11]. Career interest plays an important role in people's career choice and development. Career interest mainly affects the individual's career orientation, stimulates people's exploration and creation, and enhances people's career adaptability and stability [12]. Therefore, vocational interest is one of the influencing factors in evaluating students' learning status.

Rounds J and Tracey TJ tested the universality of Holland's hexagonal structure model [13]. Holland's career interest theory has been verified and tested for decades, the system is perfect, it conforms to logical and empirical scientific standards, and has a wide range of application examples, so it has a huge impact in the field of career interest research [14]. Combining the Holland Career Interest Inventory and the research theme of this article, compile a more adapted Occupational Interest Questionnaire. Those who meet the following "Yes" or "No" answers are scored 1 point, and those that do not meet are scored 0 points. Each student was marked with occupational orientation results according to the Vocational Interest Questionnaire scoring criteria. Quantification method: Rank the top three types from highest to lowest to produce a three-dimensional combination of answers. Realistic (R), research (I), artistic (A), social (S), enterprise (E) and traditional (C) are divided into 1, 2, 3, 4, 5 and 6, and then the total score of 6 is converted into the total score of 10. The sum of the first three occupation types obtained by each student is recorded as X, and the three times of the total score of each type is recorded as SUM. Then the quantitative formula of Career Orientation (CO) is:

$$CO(i) = \frac{10Xi}{SUM}. \tag{1}$$

B. Self-Control Index

Students are easily distracted when they are in a period of rapid youth development. They also have some distinctive characteristics of adolescents, such as being easy to accept new things and being greatly influenced by the environment. These characteristics determine that they are likely to be interfered by external factors when doing something. Unable to focus.

Students generally are unable to concentrate for a long time, which is a distinguishing characteristic of them. The deep reason for concentration is self-control. Because students are generally lack of self-control, they cannot concentrate on the teaching atmosphere created by the teacher. They are easily disturbed by intentional or unintentional attention from the outside world and cannot concentrate on the classroom.Self-control ability is an important manifestation of personal willpower and an important guarantee for the realization of personal self-development [15]. Therefore, students' self-control ability is extremely important to their daily learning. Observing the changes of students' self-control ability in the teaching process is also an important link in the implementation of the teaching process.

Self-control ability affects the development of everyone, and the research on self-control has developed rapidly in recent years. The Self-Control Questionnaire (The Self-Control Questionnaire) originally published by Brandon et al. in 1990 is actually a self-control questionnaire for healthy behaviors, in which one-quarter of the items that investigate eating habits are surveyed, and self-control is far more than just healthy behaviors [16]. Baumeister et al. proposed the theory of limited self-control and pointed out that self-control includes four areas: thought control, emotional control, impulse control, and behavioral performance [17]. In 2004, Tangney published a new self-control scale, which was tested and promoted in the United States. The test results showed that the scale has good reliability and validity [18]. Shuhua Tan and Yongyu Guo tested the self-control questionnaire on 799 Wuhan university students in 2008, and conducted confirmatory factor analysis and reliability and validity tests on the scale. The scale is composed of 19 items, with 5 levels of rating, 5 points for "completely inconsistent", 4 points for

"somewhat inconsistent", 3 points for "unsure", 2 points for "comparatively consistent", and 1 point for "completely consistent". The scale contains 5 dimensions, namely impulse control, healthy habits, resisting temptation, focusing on work, and temperance in entertainment. The lower the total score, the worse the trait self-control. T This paper will use the revised Self Control Scale to evaluate students' self-control index [19]. Quantification method: According to the total score of 95 points, the student's score is converted into a score of 10 points. The score of each student's self-control scale is recorded as *Y*, and the total score of the self-control scale is recorded as *SUM*. Then the quantitative formula of Self-Control Index (SC) is:

$$SC(i) = \frac{10Yi}{SUM}. \tag{2}$$

C. Knowledge Level

The basic level of students is not neat. The intellectual development among different individuals is unique: First, different individuals reflect different intelligence distributions, that is, each individual has a unique combination of intelligence and way, and has different intellectual strengths and weaknesses; Second, there are differences in the manner of expression and development speed of the same intelligence in different individuals [20]. Exploring the development level of students' meta-comprehension is of great significance for analyzing the differences in students' understanding level and formulating effective measures to promote the improvement of students' understanding level [21]. In the process of teaching, teachers should respect the individual differences of students, acknowledge the uniqueness of students, and pay attention to the longitudinal unique development of students.

It is an important influencing factor to treat students with a developmental perspective in the teaching process. Periodic test is a formative evaluation in the teaching process, which can quickly understand the knowledge level of students at the present stage. It is an essential quality for a teacher to adjust the teaching progress according to the students' knowledge level. Periodic test scores become a data source of Knowledge Level. Quantitative method: According to the weight of the total score of each course, the student's score is converted into a score of 10 points. The total score of each student's N courses is 100, and the credits are respectively recorded as (W1, W2..., Wn). The corresponding score of each student's course is recorded as (Z1, Z2..., Zn), and finally the total score is converted into 10 The quantified formula of Knowledge Level (KL) is:

$$KL(i) = \frac{\sum_{i=1}^{N} ZiWi}{10\sum_{i=1}^{N} Wi}. \tag{3}$$

III. Application Of DBSCAN Clustering Algorithm

A. Algorithm framework

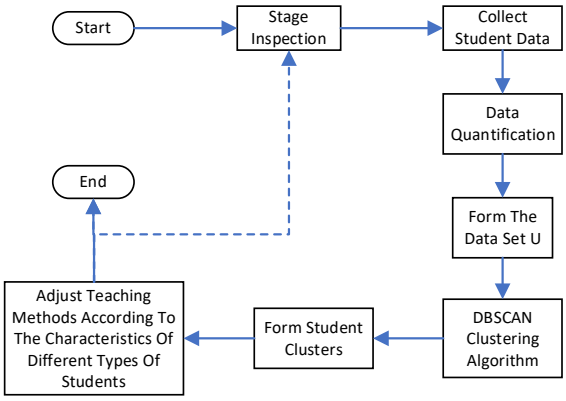


Figure I. Flow chart

As shown in Fig. 1, algorithm framework is as follows:

Step 1: Select a certain stage for stage testing, complete the Occupational Interest Questionnaire and Self-Control Scale.

Step 2: Collect student data.

Step 3: Quantify the three-dimensional data according to the quantization formula to form a data set.

Step 4: Clustering the data set by DBSCAN clustering algorithm to form student clusters.

Step 5: Adjust teaching methods according to the characteristics of different types of students.

B. DBSCAN Clustering Algorithm to Cluster Students

Details of above step 4 are as follows:

Step 1: Combine the student's Career Orientation (CO) *X*, Self-Control Index (SC) *Y*, and Knowledge Level (KL) *Z* into a three-dimensional array data source.

Step 2: Determine the radius *Eps* and the minimum number of points *MinPts*. Calculate the Euclidean distance between each point and other points, calculate the K-distance of each point, and sort the K-distance of each point in ascending order, and display the K-distance of all points in a scatter chart. Through observation, the value of K-distance corresponding to the position that changes sharply is determined as the value of radius *Eps*. The teacher determines the value of *MinPts* based on the minimum number of students with higher similarity in the class based on experience.

Step 3: Initialization: core object set Ω=∅, cluster number k=0, unvisited sample set T=D, cluster division C=∅. Calculate the distance to find all the core objects, determine whether the core object set Ω is empty, and end if it is

empty; if it is not empty, randomly select the core object o. Find the cluster sample set according to the neighborhood concept, and update the unvisited sample set T. Find whether there is a core object in the cluster sample set. If there is a core object, look for the sample set in the neighborhood of the core object, the existing samples in the cluster sample set, add the difference sample to the previous sample set, and loop until the cluster There is no core object in the sample set, and the cluster is clustered. Continue to randomly select a core object from the remaining core sample sets, return to the previous step, and loop until the core sample set is empty, the algorithm ends, and the clustering is completed. The specific algorithm-flow is shown in Algorithm 1:

Algorithm 1. DBSCAN

```
Clustering(SetOfPoints,Eps,MinPts):
    for i in SetOfPoints:
        FIND ALL MainPoint;

    for i in MainPoint:
        JOIN LIST Sequence;
        for j in Sequence:
            for k in SetOfPoints:
                if(distance < Eps) and k in MainPoint:
                    if k in Sequence:
                        continue;
                    else
                        JOIN LIST Sequence;
        for j in Sequence:
            for k in SetOfPoints:
                if(distance < Eps):
                    ADD k TO CLUSTER ThisCluster;
                    REMOVE k FROM setOfPoints;
        for j in Sequence:
            if j in MainPoint:
            REMOVE j FROM MainPoint;

    CLEAR CLUSTER ThisCluster;
    CLEAR LIST SetOfPoints;
```

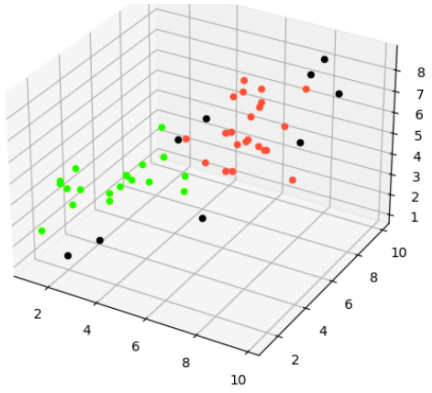
C. Experimental Data And Result Analysis

The data of the first stage of this research comes from questionnaire surveys and periodic tests conducted by 50 students of the Northwest Normal University. The data of 20 students are shown in Table 1.

It can be seen from the experimental results that 50 students are roughly divided into two categories. 22 students with medium and high indicators are marked as red, 19 students with medium and low indicators are marked as green, and the remaining 9 students are marked as green. The “noise point” is marked in black. As shown in Fig. 2.

TABLE I. EXPERTMENTAL DATA

Student	Career Orientation (CO)	Self-Control Index (SC)	Knowledge Level (KL)
Student 1	4.6	5.7	3.8
Student 2	1.2	2.8	4.1
Student 3	5.6	8.2	7.3
Student 4	3.4	6.1	5.5
Student 5	3.5	7.1	4.4
Student 6	3.8	4.6	3.9
Student 7	5.9	4.7	2.8
Student 8	4.1	5.1	4.9
Student 9	3.2	2.9	3.8
Student 10	4.2	3.5	5.5
Student 11	5.2	4.6	3.9
Student 12	3.4	3.3	4.3
Student 13	7.0	5.8	6.2
Student 14	2.4	2.3	4.4
Student 15	7.9	9.9	8.1
Student 16	6.5	7.7	6.6
Student 17	5.6	8.1	6.8
Student 18	7.6	5.8	8.2
Student 19	2.1	1.9	4.6
Student 20	6.7	5.3	6.8



FigureII. DBSCAN clustering algorithm experimental result gragh

According to the clustering results, appropriately adjust the teaching method. The red student group belongs to the students with strong learning ability and good conditions in the class. Their teaching requirements should be raised to difficult knowledge points, rather than just completing some basic knowledge learning. The green student group belongs to the students with relatively average learning ability, easy to divert attention, and lack of interest. For their teaching

requirements, they should be proficient in key knowledge and arrange interesting teaching tasks. By completing appropriate learning tasks, students' learning interest and self-confidence will increase, and their abilities will gradually improve, instead of being overly demanding and eager for success. There are two situations for students drifting outside of class: students with indicators close to full marks are students with learning ability, and students with low indicators are unable to keep up with classroom teaching and require special student attention. You can adjust the seats of these two kinds of students together to form a learning pair through advanced and backward students. The advanced students consolidate their knowledge and the backward students gradually improve to form a win-win situation.

IV. CONCLUSION

This article can be seen from the result graph. It can be seen from the result graph that using the DBSCAN clustering algorithm to classify students is more suitable for the situation in the teaching process. It can simulate the state of students in the teaching process, and use noise points to identify special students to attract the attention of teachers. However, due to the limitation of the experimental conditions, only the first phase of the experiment was carried out, which has certain limitations in terms of data.

The DBSCAN clustering algorithm is with a high similarity to the process and objects of teaching activities in terms of the amount of data which can be carried, unsupervised clustering, and noise analysis. By using the DBSCAN clustering algorithm to classify students in the class, the result graph is generated to show the learning status of the students in the class, so that teachers can quickly grasp the situation of the class.

REFERENCES

[1] Y. Ma and H. Ye, "Contemporary development of Vygotsky's recent development zone thought," Psychological Development and Education, no.02, pp. 89-93, 2004.

[2] Z. Xiang, "The inner connection between interest and teaching effect," Hubei Primary and Secondary School Laboratory, vol.4, no.06, pp.34-35,1994.

[3] F. Guo, "Cultivating students' character——improving English teaching effect," China Science Education Innovation Guide, no.25, pp.130-130,2012.

[4] L. Xiao, T. Hu, Y. Liu, Z. Lin, Y. Wu, "Research on student status analysis system based on hybrid clustering algorithm," Application Research of Computers, vol.29, no.06,pp.174-177, 2012.

[5] Abualhaj B, Weng G, Ong M, Attarwala AA, Molina F, K Büsing K, Glatting G, "Comparison of five cluster validity indices performance in brain [18F]FET-PET image segmentation using k-means," Medical Physics, vol.44, no.01,pp.209-219, 2017.

[6] L. Lu, M. Yu, "An analysis method of student learning status based on outlier detection," Computer and Modernization, no.247,pp.35-40, 2016.

[7] X. Wu, "Research on the Diversified Application of Computer Technology in Education," Science and Technology Innovation Herald, no.08,pp.218-219, 2020.

[8] K. Liang, "On Personality and Career Choice," Education and Career, no.14, pp.59-60,2006.

[9] F. Li, "The Enlightenment of the Different Types of Occupational Interests of Students on Education," Global Education Outlook, vol.40,no.01,pp.69-72, 2020.

[10] F. Wang, Y. Shi, "Survey and Research on the Influence of Students' Personality on Vocational Interests in Higher Vocational Colleges," Journal of Changchun University of Science and Technology (Higher Education Edition) , vol.4, no.06, pp.11-12,2009.

[11] J. Shen, "Review of Holland's Vocational Interest Theory and Its Application," Vocational Education Research, vol.07, pp.9-10,2010.

[12] L. Wang, "Holland's career choice theory and its practical application," Quest, no.07, pp.160-161,2009.

[13] Rounds J, Tracey TJ, "Prediger's dimensional representation of Holland's RIASEC circumplex," Journal of Applied Psychology, 1993.

[14] L. Bai, "Introduction and Review of Holland's Occupational Interest Theory," Advances in Psychological Science, pp.27-31,1996.

[15] Z. Ji, "Self-knowledge, Self-Confidence, Self-Control: A Preliminary Study of the 'Three-Self Education' Cultivation Model for Postgraduates of Nationalities Universities," Degrees and Postgraduate Education, no.04, pp.31-35,2009.

[16] Tangney J P , Baumeister R F , Boone A L, "High self-control predicts good adjustment, less pathology, better grades, and interpersonal success," Journal of Personality, vol.72, no.2, pp.271-324,2004.

[17] S. Tan, Y. Guo, "Theoretical hypothesis of limited self-control and related research," Chinese Journal of Clinical Psychology, vol.16, no.3, pp.309-311,2008.

[18] S. Tan, Y. Guo, "Revision of the Self-Control Scale for College Students," Chinese Journal of Clinical Psychology, vol.16,no.5, pp.468-470,2008.

[19] X. Xu, "Self-control deficiency and its mechanism in college students with mobile phone addiction," Southwest University , 2014.

[20] L. Huo, Q. Zhao, "The Evaluation View of Multiple Intelligence Theory and Its Enlightenment to the Evaluation of Student Development," Comparative Education Research, vol.4, pp.45-50,2005.

[21] H. Chen, Q. Dong, "A Study on the Relationship between Middle School Students' Comprehension Level and Meta-understanding Knowledge," Psychological Development and Education, no.03, pp.7-14,1992.