Information Classification Algorithm based on Project-based Learning Data-driven and Stochastic Grid

Xiaomei Qin, Wenlan Zhang*
School of Education, Shaanxi Normal University, Xi'an Shaanxi 710062, China

***Abstract*—The adaptive partitioning algorithm of information set in simulation laboratory based on project-based learning data-driven and random grid is studied to effectively preprocess the information set and improve the adaptive partitioning effect of the information set. Using the improved fuzzy C-means clustering algorithm driven by project-based learning data, the fuzzy partition of information set in simulation laboratory is carried out to complete preprocessing of information set; The pre-processing information set space is roughly divided by the grid partitioning algorithm based on the data histogram; A random mesh generation algorithm based on uniformity is used to finely divide the coarse mesh cells; Taking the representative points of grid cells as the clustering center, the pre-processing information set is clustered by the density peak clustering algorithm to complete the adaptive partitioning of the information set in simulation laboratory. Experimental results show that this algorithm can effectively preprocess and adaptively partition the information set of simulation laboratory; For different dimension information sets, the evaluation index values of Rand index, Purity, standard mutual information, interval and Dunn index of the algorithm are all high, and the evaluation index values of compactness and Davidson's banding index are all low, so the algorithm has a high accuracy of adaptive partitioning of information sets.**

***Keywords*—Adaptive partitioning; data driven; information set; project-based learning; random grid; simulation laboratory**

I. Introduction

Simulation experiments exist in the simulation laboratory, and the experimental environment, objects and equipment are provided by the simulation laboratory [1]. The simulation laboratory is a kind of simulation experiment environment supported by computer software and hardware technology and realized by software development tools [2]. By developing a series of simulation experiment components to simulate and reproduce the experimental environment, experimental equipment and experimental process, the experimenter can get rid of the bondage of the actual experimental conditions, feel the experimental information interactively, and realize the experimental process in a near real way under more convenient and fast conditions [3]. In the simulation laboratory, the experimental objects and equipment are either reproduced or simulated vividly, and the experimental process is completely controlled by the experimenter [4]. The simulation laboratory uses the powerful computing processing ability of the computer, with the help of graphics / images, simulation and virtual reality technologies, and has rich interface information, friendly interaction ability and powerful data processing function. In addition, it is well compatible with various external devices, multi-media and Internet, forming a wonderful simulation experiment world. Project-based learning is a new teaching mode, which mainly focuses on the core concepts and principles of the discipline, and emphasizes that learners can solve practical problems, participate in some exploratory activities and other meaningful learning tasks [5]. In this process, learners learn independently and construct the meaning of the learned knowledge through practical operation, and apply it in the simulation laboratory. It can effectively improve learners' interest in simulation experiments and learning quality. Once the project-based learning simulation laboratory is established, it is a shared resource. Although it is not limited by the site and time, it greatly saves material resources, but it will generate a large number of project-based learning data and increase the difficulty of data search. The best way to solve this problem is to study a partitioning algorithm, which divides all information into several categories through the partitioning algorithm to facilitate information search. For example, Sutagundar, A., et al. proposed to use the function of sensor cloud and fog calculation to divide in a better way and minimize the delay problem. Using random forest classifier and genetic algorithm to divide the information and introducing Agent paradigm, not only saved the energy of physical sensor nodes, but also could quickly analyze and divide the information on the fog server. The results show that the algorithm has better effect in partition accuracy, delay and energy consumption [6]; Flisar, J., et al. used DBpedia ontology knowledge base to divide information. This algorithm can effectively divide information sets [7], but it cannot adapt to all data distribution well. The partitioning algorithm boundary is easily discarded as noise points, resulting in low partitioning accuracy. In order to achieve efficient, high-precision and high-speed information partition, Zheng, T., et al. designed a partitioning algorithm with low power consumption and high performance. In the training phase, a laser radar was simulated to collect the laser radar information, and then the designed neural network was trained using the information and the corresponding tags. In the testing phase, the new information was first divided into simple units using the range transformation watershed method, then, the trained neural network was used to divide the information. The average recall rate of the information partitioning of the algorithm was 0.965 and the average precision rate was 0.943 [8], but its complexity was high and it was difficult to meet the real-time requirements, and there were limitations in the processing of

*Corresponding Author.

multidimensional information sets. The algorithm is susceptible to noise and the effect of boundary processing is not ideal. Data driving can be broadly defined as using the online or offline data of the system to realize various expected functions of the system such as data-based preprocessing, evaluation, scheduling, monitoring, diagnosis, decision-making and partitioning algorithm [9]. The random grid density clustering algorithm is insensitive to the input data and can adapt to different data distributions. Moreover, the algorithm has low complexity and fast calculation speed. It is suitable for partitioning different dimensional information and is not sensitive to the impact of noise [10].

In order to improve the precision of adaptive partitioning of information set, an adaptive partitioning algorithm of information set in simulation laboratory based on project-based learning data-driven and random grid is studied. First of all, the paper uses the fuzzy C-means clustering algorithm to carry out the fuzzy division of information, and preliminarily classifies the information. On this basis, the paper further uses the data histogram grid division algorithm to further roughly divide the information space, strengthen the classification of information, and avoid the problem of insufficient classification caused by the strong correlation between information; Then, the information is further clustered with the density peak clustering algorithm, which strengthens the aggregation of similar information and completes the information division. Finally, the experimental results show that this research method has the ability of adaptive partition, high precision of partition, and good overall application.

II. Adaptive Partitioning Algorithm of Information Set in Simulation Laboratory

A. *Information Set Preprocessing of Simulation Laboratory by using the Improved Fuzzy C-means Clustering Algorithm based on Project-based Learning Data-driven*

In the simulation laboratory, the project-based learning method is applied to carry out simulation experiments, so that students can explore the selected simulation experiment projects according to their own interests and learning needs. This learning activity emphasizes the students' hands-on operation and comprehensive application of various knowledge achievements, which can improve the learning quality of each simulation experiment project stage and make students more interested in the learning process of simulation experiments. In the process of simulation experiments using project-based learning, a large number of project-based learning data will be generated in the information of the simulation laboratory. In order to further improve the learning quality of each project stage in the simulation experiment process, it is necessary to self-adapt the project-based learning data in the information set of the simulation laboratory [11], accelerate the efficiency of students viewing the project-based learning data in the information set of the simulation laboratory, and facilitate the management of the information set of the simulation laboratory. In order to improve the adaptive partitioning effect of the project-based learning data in the information set of the simulation laboratory [12], it is necessary to obtain the fuzzy version of the project-based learning data in the original information set of the simulation laboratory (composed of fuzzy attributes and fuzzy partitions). Fuzzy C-means clustering algorithm (FCM) generates fuzzy partitions by defining the value of fuzzy membership function ( ), which can solve the problem of hard boundary value caused by sharp partitions and protect the item learning data in the original information set. In addition, the fuzzy partition has obvious semantic relevance, which can well solve the inherent uncertainty of numerical data in the project-based learning data in the information set of the simulation laboratory.

An expression called fuzzy entropy is used as the cost function of the objective function of FCM algorithm. The definition of fuzzy entropy is roughly the same as that of information entropy, which is more suitable for fuzzy clustering analysis [13]. The function of fuzzy entropy can be defined as:

$$E(x) = -\sum_{i=1}^{c}\sum_{j=1}^{n} m\mu_{ij} \ln m\mu_{ij} \tag{1}$$

Wherein, the project-based learning data in the information set of simulation laboratory is x; The fuzzy membership degree of the i-th attribute of the project-based learning data in the i-th information set of simulation laboratory is μ_{ij}; m is a weighted index, whose physical meaning is the fuzziness degree constant; The number of project-based learning data in the simulation laboratory information set is c; The number of fuzzy attributes of the project-based learning data in the information set of the simulation laboratory is n.

The objective function of FCM can be defined as:

$$\min J(\mu, v) = \sum_{i=1}^{c}\sum_{j=1}^{n} m\mu_{ij} d_{ij}^{2}$$

$$s.t. \sum_{i=1}^{c} \mu_{ij} = 1 \tag{2}$$

Where, the objective function of the sum of error squares is $\min J(\mu, v)$, and its value reflects the degree of compactness within the class under a certain difference definition [14]. The smaller the value of $\min J(\mu, v)$ is, the tighter the clustering is; The clustering center of project-based learning data in the information set of simulation laboratory is v; The Euclidean distance is d_{ij}^{2}.

Taking information entropy as the cost function, the minimum error square sum objective function of FCM algorithm is used to solve the problem of fuzzy partition of project-based learning data in the information set of simulation laboratory. The objective function of the improved FCM algorithm is:

$$L(\mu, v, \alpha) = \min J(\mu, v) + \alpha E(x) \tag{3}$$

Where the Lagrange multiplier is α.

Substituting formula (1) and formula (2) in formula (3) can obtain:

$$L(\mu,v,\alpha)=\sum_{i=1}^{c}\sum_{j=1}^{n}m\mu_{ij}d_{ij}^{2}-\sum_{i=1}^{c}\sum_{j=1}^{n}\alpha_j m\mu_{ij}\ln m\mu_{ij}^{m} \tag{4}$$

The Lagrange multiplier of the j-th attribute of the project-based learning data in the information set of the simulation laboratory is α_j.

If $\frac{\partial L}{\partial \mu_{ij}}=0$ in formula (4), the following can be obtained:

$$\mu_{ij}=\exp\left(\frac{d_{ij}^{2}}{m\alpha_j}-\frac{1}{m}\right) \tag{5}$$

Since $\sum_{i=1}^{c}\mu_{ij}=1$, the finishing formula (5) can be obtained:

$$\alpha_j=\frac{1}{\sum_{k=1}^{c}d_{kj}^{2}} \tag{6}$$

Wherein the attribute number of the project-based learning data in the information set of simulation laboratory is k.

Substituting formula (5) and formula (6) into formula (4), it can get:

$$L(\mu,v,\alpha)=\frac{m\mu_{ij}}{\sum_{k=1}^{c}d_{kj}^{2}} \tag{7}$$

It can be seen from formula (7) that in order to obtain the optimal value of $\min J(\mu,v)$, it is also necessary to process the project-based learning data in the information set of simulation laboratory [15]. By analyzing the physical meaning of $\frac{1}{\sum_{k=1}^{c}d_{kj}^{2}}$, it can be seen that in fact it should be the distribution characteristics of project-based learning data in the information set of the simulation laboratory, representing a distribution characteristic of project-based learning data in the data space [16]. That is to say, taking information entropy as the cost function of the objective function, the objective function of the FCM algorithm needs to obtain the minimum value. In addition to the membership degree [17], the actual distribution characteristics of the project-based learning data in the data space in each cluster must also be considered. For the convenience of later expression and calculation, $\delta_j=\frac{1}{\sum_{k=1}^{c}d_{kj}^{2}}$ is specially used, that is, δ_j is used to represent this distribution characteristic.

According to the above analysis, the objective function is adjusted accordingly to obtain:

$$J(\mu,v,\delta)=\sum_{i=1}^{c}\sum_{j=1}^{n}m\delta_j m\mu_{ij}d_{ij}^{2}$$

$$s.t.\sum_{i=1}^{c}\mu_{ij}=1 \tag{8}$$

Then, according to the Lagrange function method, it can get:

$$J(\mu,v,\delta,\alpha)=\min\sum_{i=1}^{c}\sum_{j=1}^{n}m\delta_j m\mu_{ij}d_{ij}^{2}+\sum_{j}^{n}\alpha_j\left(1-\sum_{i=1}^{c}\mu_{ij}\right) \tag{9}$$

Let $\frac{\partial L}{\partial v_i}=0$, then:

$$v_i=\frac{\sum_{j=1}^{n}m\delta_j m\mu_{ij}x_j}{\sum_{j=1}^{n}m\delta_j m\mu_{ij}} \tag{10}$$

Wherein the project-based learning data in the information set of simulation laboratory of the j-th attribute is x_j; The clustering center of the project-based learning data in the i-th simulation laboratory information set is v_i.

Let $\frac{\partial L}{\partial \mu_{ij}}=0$, then:

$$\mu_{ij}=\left(\frac{\alpha_j}{m\delta_j d_{ij}^{2}}\right)^{\frac{1}{m-1}} \tag{11}$$

The fuzzy partition of the project-based learning data in the information set of the simulation laboratory is generated by iteratively optimizing the objective function formula (8) by updating the values of the membership function μ_{ij} and the cluster center v_i.

In order to improve the fuzzy partition effect of the project-based learning data in the information set of the simulation laboratory, the fuzzy membership function of the FCM algorithm is determined by using the project-based learning data-driven method. The generation process of fuzzy partition of project-based learning data in the information set of simulation laboratory based on project-based learning data driven FCM is as follows:

It is defined that the project-based learning data in the information set of the simulation laboratory is $D=\{x_1,x_2,\cdots x_c\}$, $x_1,x_2,\cdots x_c$ is the data of different numerical attributes, and D is the set of Boolean attributes and numerical attributes; The numeric attribute set is defined as $A=\{q_1,q_2,\cdots,q_r\}$; The fuzzy partition is defined as $P=\{P_1,P_2,\cdots,P_r\}$, the fuzzy partition is the result of clustering the numerical attribute set A by FCM algorithm, and $P_r=\{f_1,f_2,\cdots,f_w\}$ is the fuzzy clustering partition set of the numerical attribute q_r.

The improved FCM algorithm is used to cluster the data in the numerical attribute set of project-based learning data in each information set of simulation laboratory, and then the corresponding fuzzy partition is obtained. Each numerical attribute data has its own unique fuzzy membership function. This process is repeated several times until the fuzzy partition of all numeric attribute values is obtained.

The whole preprocessing process of the project-based learning data in the information set of the simulation laboratory includes two steps: the first step is to generate fuzzy partitions for the values of the numerical attributes of the project-based learning data in each information set of the simulation laboratory through the improved FCM algorithm; The second step is to further process the project-based learning data in the original information set of simulation laboratory, and then obtain its fuzzy Version (composed of fuzzy attributes and fuzzy partitions). Pre definition: the Boolean attribute value data set is $B=\{b_1,b_2,\cdots,b_{m'}\}$; The definition attribute set is $\hat{A}=B\cup A$. The fuzzy version generation process of the project-based learning data in the original information set of simulation laboratory is as follows: scan the project-based learning data in the original information set of simulation laboratory, classify the data attribute values, put the project-based learning data belonging to the Boolean attribute in $B=\{b_1,b_2,\cdots,b_{m'}\}$, and put the project-based learning data belonging to the numerical attribute in $A=\{q_1,q_2,\cdots,q_r\}$, and then perform classification calculation. When an item learning data is a Boolean attribute, its fuzzy membership function $\mu=1$ or $\mu=0$ can easily divide the fuzzy partition, and then obtain the fuzzy version of the Boolean attribute value [18]; When a project-based learning data is a numerical attribute, each numerical attribute in the project-based learning data D can be converted into a fuzzy record according to the fuzzy partition P_r. Each fuzzy record contains the fuzzy attribute of the project-based learning data and the corresponding fuzzy membership function.

Thus, after selecting the data attributes in the first step, an intermediate version of the project-based learning data D_1 in the information set of simulation laboratory can be generated; Then, D_1 is updated iteratively [19], until all the attribute data in the attribute set $\hat{A}$ are processed, and then the fuzzy version of the original project-based learning data will be obtained. Therefore, by applying the FCM preprocessing technology driven by project-based learning data, any project-based learning data in the original information set of simulation laboratory with Boolean and numerical attributes can be transformed into a fuzzy set $Q=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_N\}$ with fuzzy records and fuzzy attributes.

B. Adaptive Partitioning of Information Set in Simulation Laboratory based on Random Grid

1) Random grid partitioning of information set in simulation laboratory: The density peak clustering algorithm based on random grid partitioning is used to adaptively divide the fuzzy set $Q=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_N\}$ of the project-based learning data in the information set of simulation laboratory obtained in Section 2.1.

In order to better adapt to the best partitioning of the fuzzy sets of project-based learning data in different situations, first, the fuzzy sets of project-based learning data are roughly divided by the data histogram, and then the random grid is divided by the grid uniformity.

In order to adapt to the adaptive partitioning environment of fuzzy sets of project-based learning data, overcome the shortcomings of current grid division, and improve the efficiency and reliability of grid cell uniformity [20], firstly, rough partitioning of fuzzy sets of project-based learning data is carried out by using histograms. The steps are as follows:

Step 1: read in the fuzzy set $Q=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_N\}$ of the project-based learning data and normalize it. The normalization formula is:

$$\tilde{x}=\frac{\hat{x}-\hat{x}_{\min}}{\hat{x}_{\max}-\hat{x}_{\min}} \tag{12}$$

Wherein, the minimum and maximum values in the fuzzy set of project-based learning data in the information set of the simulation laboratory are $\hat{x}_{\min}$ and $\hat{x}_{\max}$; The item learning data in the fuzzy set after normalization processing is $\tilde{x}$.

Step 2: scan the fuzzy set Q of item learning data once, and draw the data histogram on each dimension within Q.

Step 3: according to the corresponding data histogram of each dimension, determine the high-density area and the low-density area of the fuzzy set of project-based learning data in the information set of the simulation laboratory, and divide the grid cells of the high-density area into fine ones and the grid cells of the low-density area into coarse ones. The specific operation is to use the number of data points contained in a grid to be equal to a given threshold β to divide, and the value of threshold β is related to the size of the entire fuzzy set Q.

Although the rough grid partitioning based on data histogram divides the regions with different densities of the project-based learning data fuzzy set in different scales, the distribution of the project-based learning data of the grid cells with the same density may still be very different, so the boundary of the class cannot be effectively found and the class with arbitrary shape cannot be effectively found. Next, the random grid cells are finely divided by the uniformity index. The steps are as follows:

Step 1: calculate the uniformity of each random grid element obtained by rough division, and the formula is as follows:

$$S_l\left(u\right)=1-\left\{\max\left(\frac{\phi_{i'}\left(u\right)-\eta_{i'}\left(u\right)}{\eta_{i'}\left(u\right)}\right)\right\} \tag{13}$$

Where $\eta_{i'}\left(u\right)$ is used to represent the overall standard deviation of the i' -dimension of the random grid unit u; $\phi_{i'}\left(u\right)$ represents the standard deviation of the i' -dimensional project-based learning data samples of the random grid cell u, and $S_l\left(u\right)$ represents the uniformity of the l-dimensional grid cell u; $i'=1,2,\cdots,l$; The closer the value of $S_l\left(u\right)$ is to 1, the higher the uniformity of the random grid cells in the fuzzy set of item learning data is.

If $S_l\left(u\right)$ satisfies the given threshold, i.e. it is a uniform mesh, it will not be divided; If the uniformity is less than the given threshold, i.e. non-uniform mesh, go to step 2.

Step 2: for each non-uniform grid, divide the random grid cell into two new grids of equal size along the worst dimension of $S_l\left(u\right)$, and judge whether the two new grids meet the stop criteria. If not, further divide them by the same method until the stop criteria are met.

Step 3: stop criteria. ① The grid element is a uniform grid; ② The grid cell is an empty grid; ③ The number of project-based learning data points in the grid is less than a given threshold.

2) Automatic selection of clustering center of information set in simulation laboratory: It is assumed that the center point of the grid unit u after random grid partitioning is $g_u=\left(g_u^1,g_u^2,\cdots,g_u^l\right)$ and $u=1,2,\cdots,K$; K is the total number of meshes; Where g_u^l represents the coordinate of the center point of the grid cell u in the l-dimension, so the grid cell G_u can be expressed as:

$$G=\left[\left(g_u^1-\frac{side}{2},g_u^1+\frac{side}{2}\right),\right.$$
$$\left.\left(g_u^2-\frac{side}{2},g_u^2+\frac{side}{2}\right),\cdots,\frac{de}{2}\right),\cdots,\left(g_u^l-\frac{side}{2},g_u^l+\frac{side}{2}\right)\right] \tag{14}$$

Wherein the side length of the grid cell is side.

The set of points in the grid cell u is $Y=\left\{y_1,y_2,\cdots,y_{h_u}\right\}$, and h_u is the total number of project-based learning data points in the grid cell u, then the representative points of the grid cell are:

$$Y_u=\frac{\sum_{\hat{x}_i\in G_u}\hat{x}_i}{\lambda h_u} \tag{15}$$

Where the constant is λ; The item learning data in the fuzzy set of the i-th information set in simulation laboratory after normalization is $\hat{x}_i$.

The local density of the grid cell's representative point Y_u is the number of project-based learning data points in the grid cell u, and the number of points in the grid cell u is:

$$h_u=\sum_{i=1}^{N}f'\left(\hat{x}_i,G_u\right) \tag{16}$$

Where, the function is $f'\left(\cdot\right)$, $f'\left(\hat{x}_i,G_u\right)=\begin{cases}1 & g_u^l-\frac{side}{2}\le\hat{x}_i^l<g_u^l+\frac{side}{2}\\ 0 & other\end{cases}$, so the local density of Y_u is $\rho_u=h_u$.

The nearest distance between the grid cell's representative point Y_u and the higher density representative point Y_o is taken as the distance value of the grid cell's representative point Y_u , which is recorded as d'_u , and the formula is as follows:

$$d'_u=\min_{o:\rho_o>\rho_u}\left(D'_{ou}\right) \tag{17}$$

Wherein d'_u is the distance between the grid cell's representative point Y_u and the grid cell's representative point Y_o .

An improved adaptive method is designed to complete the automatic selection of cluster centers when the information set of simulation laboratory is adaptively divided, and the exact number of cluster centers is selected without manual intervention, so as to improve the accuracy of cluster center selection. The determination function is:

$$\rho_{C_{\hat{i}}}-\sigma\left(\rho_{\hat{i}}\right)\ge 0 \tag{18}$$

$$\frac{\xi_{C_{\hat{i}}}-E'\left(\xi_{\hat{i}}\right)}{2}\ge\varpi\left(\xi_{\hat{i}}\right) \tag{19}$$

Wherein the density of the representative point $C_{\hat{i}}$ of the grid cell of the $\hat{i}$ -th cluster center is $\rho_{C_{\hat{i}}}$; The mean value of the representative point density of all grid cells is $\sigma\left(\rho_{\hat{i}}\right)$; The minimum distance between the representative point of grid cell and the representative point of cluster center in the same cluster is $\xi_{C_{\hat{i}}}$; The expectation of all $\xi_{\hat{i}}$ is $E'\left(\xi_{\hat{i}}\right)$. Formula (18) indicates that the local density value of the representative points of the grid cell is greater than the average value of the local density of all the representative points in the grid cell. This determination method satisfies the condition that the clustering centers of the project-based learning data in the information set of the simulation laboratory are often distributed in the relatively high density area in the density peak algorithm. The determination method of formula (19) satisfies the condition that the relative distance between the cluster centers is relatively long. Therefore, when the grid cell's representative point object meets the above two formula conditions, the grid cell's representative point is selected as the cluster center.

3) Classification of project-based learning data points in the information set of simulation laboratory: The nearest neighbor algorithm in the density peak clustering algorithm is used to classify the item learning data points in the fuzzy set of the remaining information set of simulation laboratory. After the selection of the representative points of the cluster center is completed, the remaining non cluster center representative points are classified into the class of the representative points closest to them and with local density greater than the point in $\rho_{\hat{i}}$ -descending order, and the data points in the project-based learning data in the fuzzy set of the original information set in simulation laboratory are assigned to the class of the representative points of the grid cells.

When the project-based learning data in the fuzzy set of the information set of the simulation laboratory is adaptively divided, the object of the boundary point is the representative point of the grid cell. Firstly, the set of boundary grid cell's representative points in the current cluster is calculated according to the density parameter θ_C, to find the grid cell's representative points with the highest density in the boundary point set, and take the density of the representative points as the threshold to divide the core representative points and noise points, so as to reserve the representative points with the density greater than or equal to the density threshold as the core representative points in the cluster; The noise representative points in the current category that are smaller than the density threshold are removed, and the project-based learning data points in the grid where the noise representative points are located are also removed.

4) Adaptive partitioning process of information set in simulation laboratory: The specific steps of the adaptive partitioning of the information set in simulation laboratory are as follows:

Step 1: use the project-based learning data in Section 2.1 to drive FCM, preprocess the project-based learning data in the information set of the simulation laboratory, and obtain the fuzzy set of the project-based learning data;

Step 2: normalize the item learning data in the fuzzy set;

Step 3: roughly mesh the fuzzy set of project-based learning data according to the data histogram;

Step 4: perform random mesh refinement on the coarse divided mesh according to the uniformity to obtain several disjoint mesh elements;

Step 5: map the project-based learning data points to the corresponding grid cells, obtain the representative points of each grid cell from formulas (14) and (15), and count the number of project-based learning data points contained in each grid cell;

Step 6: calculate the local density ρ_u of the representative point Y_u of the grid cell according to formula (16);

Step 7: arrange the grid cell's representative points in reverse order according to $\rho_{\hat{i}}$, and calculate the high density distance d'_u of each grid cell's representative point according to formula (17);

Step 8: adaptively determine the representative point of the cluster center by formula (18) and formula (19), classify it into the class of the grid representative point with the shortest distance and the local density greater than the point according to the p-descending order, and classify all data points in the project-based learning data in the original information set of simulation laboratory into the class of the grid representative point;

Step 9: calculate the boundary point set of the current class from the density parameter θ_C, select the representative point with the highest density in the boundary point set, and use the density of the point as the threshold for dividing the core representative points and noise points of the current class, and eliminate the representative points in the current class that are smaller than the density threshold and other item learning data points in the grid cell where the representative points are located;

Step 10: return the final clustering result, that is, complete the adaptive partitioning of the simulation laboratory information set.

III. Experimental Analysis

Taking the sensor simulation laboratory of a university as the experimental object, the simulation laboratory realizes semi-automatic interactive control through animation, video or virtual reality technology. The structure of the simulation laboratory is shown in Fig. 1.

The simulation laboratory integrates the resources such as computers, instruments and equipment, tested points and their data into the network for sharing, and realizes the functions of remote or remote testing, control, data acquisition, fault monitoring and on-site monitoring. The networked simulation instrument can obtain the measurement information from any place and at any time.

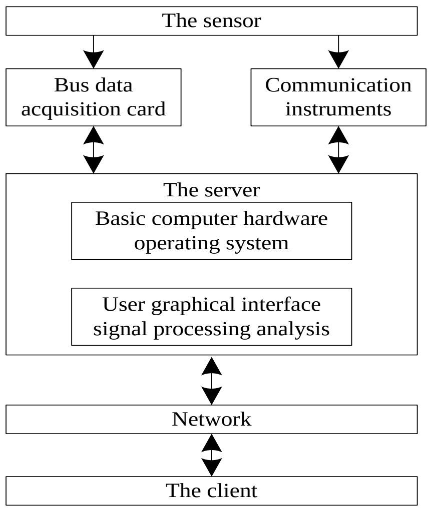


Fig. 1. Structure Diagram of Simulation Laboratory.

In the process of sensor simulation experiments in the project-based learning mode conducted by the university students in the simulation laboratory, tens of thousands of project-based learning data of different dimensions are generated, forming a high-dimensional information set in simulation laboratory and a low-dimensional information set in simulation experiment. Both information sets contain four types of project-based learning data, namely, carrier data, modulation data, excitation data and vibration data. Using the algorithm in this paper, the information sets of two simulation laboratories are divided adaptively, which proves that the algorithm in this paper has a good adaptive partitioning effect.

Taking the information set of low dimensional simulation experiment as an example, the information set is fuzzy partitioned by the algorithm in this paper, and the results of fuzzy partitioning are shown in Fig. 2.

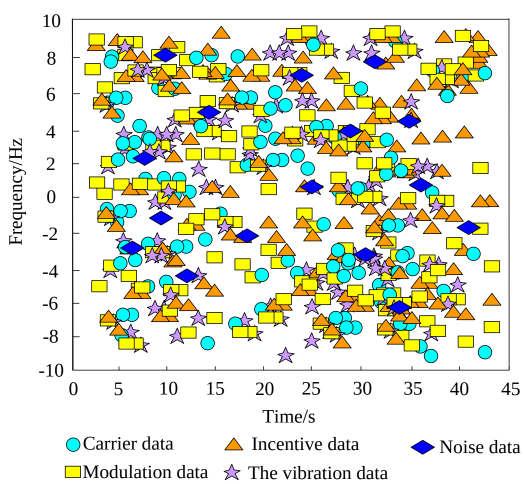


Fig. 2. Project-based Learning Data in the Original Low-dimensional Simulation Laboratory Information Set.

Using the quantitative attribute Income, the effect of fuzzy partition of the project-based learning data in the information of the low-dimensional simulation laboratory is analyzed. The value range of the Income attribute is $Income \in [0,9]$ (the unit is thousands). The number of fuzzy partitions of project-based learning data in the information set of simulation laboratory is defined as 4, which are "Around 1", "Around 3", "Around 5", "Around 7"; The analysis results of fuzzy partition effect of the algorithm in this paper are shown in Fig. 3.

According to Fig. 3, after preprocessing the project-based learning data in the original low-dimensional information set of simulation laboratory using the algorithm in this paper, the fuzzy partitions obtained are related to each other, that is, they have strong semantic relevance, and the curves of each fuzzy partition are relatively smooth, which solves the problem of "sharp partition", protects the boundary value, and thus protects the project-based learning data. The Income value corresponding to the highest value of the fuzzy membership degree of each fuzzy partition corresponds to the set Around value, which indicates that the algorithm in this paper has better fuzzy partition effect of the information set in simulation experiment, that is, better preprocessing effect of the information set.

The algorithm in this paper is used to adaptively partition the project-based learning data in the low-dimensional information set of simulation experiment. The adaptive partitioning results are shown in Fig. 4.

According to Fig. 4, the algorithm in this paper can effectively divide the project-based learning data in the information set of the simulation laboratory. After the rough division, the description between the project-based learning data is relatively fuzzy, and the noise points are not distinguished. After the random grid refinement, the boundary area between the project-based learning data is further refined, and the noise point data is effectively divided into isolated grid cells; According to the grid cells finely divided by random grid, four kinds of project-based learning data are obtained through clustering processing, and the noise data scattered outside the project-based learning data is better removed. Experimental results show that the proposed algorithm can effectively partition the information set of simulation laboratory.

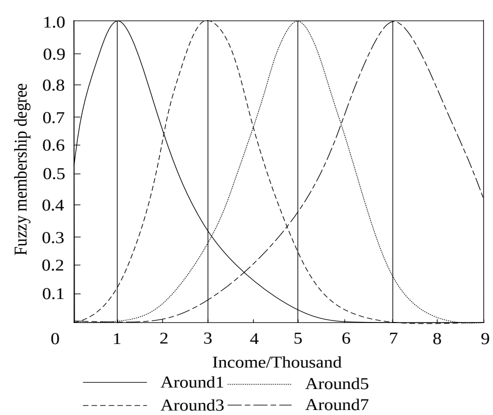
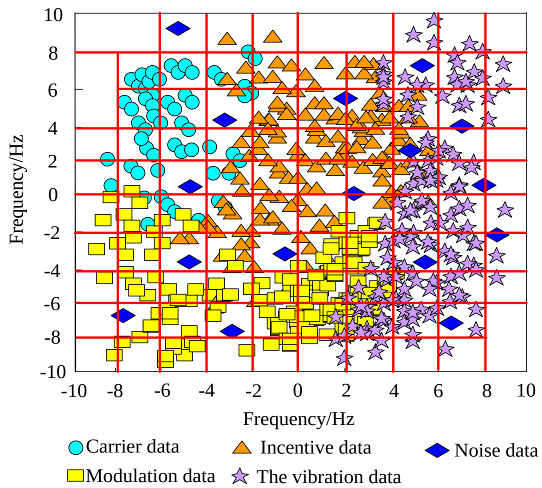
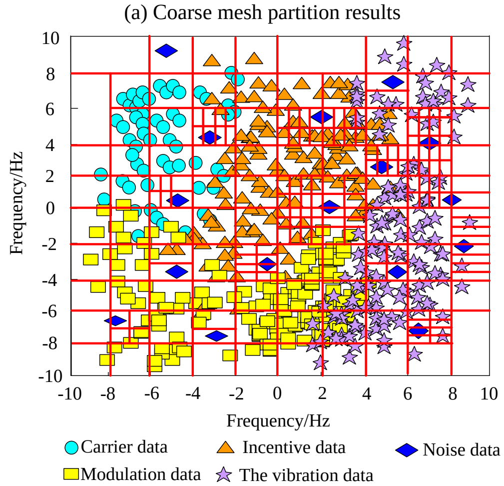


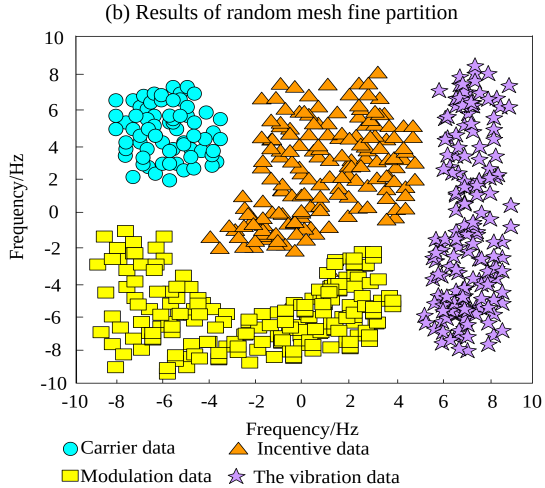
Fig. 3. Fuzzy Partition Effect.



(a) Coarse mesh partition results



(b) Results of random mesh fine partition



(c) Adaptive clustering results

Fig. 4. Adaptive Partitioning Results of Project-based Learning Data in Low-dimensional Simulation Experiment Information Set.

The external evaluation indexes Rand index, Purity and normalized mutual information (NMI), internal evaluation indexes compactness (CP), separation (SP), and relative evaluation indexes including Davies bouldin index (DBI) and Dunn validity index (DVI) are used as evaluation indexes. The higher the value of Rand index is used to measure the percentage of correct division, the higher the adaptive partitioning accuracy is, the value range is [0, 1]; Purity represents the purity of the partitioning result, and the value range is [0%, 100%]. The higher the value is, the better the partitioning result is; NMI is to evaluate the consistency between the clustering results and the real categories through the mutual information between the clustering results and the real categories. The value is between [0, 1]. The higher the value is, the better the classification effect is; The lower CP means that the closer the clustering distance within the class is, the better the compactness is, and the value range is [0, 1]; The higher the larger of SP is, the farther the cluster distance and the distance between clusters are, and the value range is [0, 1]; The closer the intra class distance is, the farther the inter class distance is, the better the partitioning effect is; The smaller DBI means the smaller the intra class distance and the larger the inter class distance, the better the partitioning effect is, and the value range is [0, 1]; The larger the DVI means the greater the distance between classes and the smaller the distance between classes, and the better the partitioning effect. The value range is [0, 1]; The effect of adaptive partitioning of information set in simulation laboratory by the algorithm in this paper is analyzed. For high-dimensional information set of simulation experiment and low-dimensional information set of simulation experiment, the test results of adaptive partitioning of information set in simulation experiment by the algorithm in this paper are shown in Table I.

According to Table I, the RAND index of the algorithm in this paper is relatively high when adaptively dividing the information sets of low-dimensional and high-dimensional simulation laboratories, which is close to 1, indicating that the algorithm in this paper adaptively divides the information sets of different dimensions with high accuracy; The purity is also high, which is close to 100%, and the NMI is high, which is close to 1. This shows that the results of adaptive partitioning of different dimension information sets in this algorithm are very similar to the actual results; SP and DVI are both large, close to 1, CP and DBI are both small, and close to 0, which indicates that the algorithm in this paper adaptively partitions the information set with a large inter class distance and a small intra class distance, and has a better adaptive partitioning effect; Comprehensive analysis shows that for different dimension simulation laboratory information sets, the algorithm in this paper can accurately and adaptively partition the information sets, and has better adaptive partitioning effect of the information sets.

In order to further verify the performance of the algorithm studied in this paper, the algorithm in the literature [6] and the algorithm in the literature [7] introduced in the introduction are used as the comparison method to test the RAND index results in different dimensions, and the test is repeated five times. The results are shown in Table II.

It can be seen from Table II that under different dimensions, the lowest RAND index of this method is 0.96, close to 1. The RAND index of literature [6] method and literature [7] method is 0.89 and 0.88 respectively, which is far lower than that of this method. This shows that the adaptive division accuracy of this method is the highest and has certain applicability.

TABLE I. TEST RESULTS OF THE PROPOSED ALGORITHM FOR ADAPTIVE PARTITIONING OF SIMULATION EXPERIMENTAL INFORMATION SETS IN DIFFERENT DIMENSIONS

The evaluation index	**Low-dimensional simulation laboratory information set**	**High dimensional simulation laboratory information set**
Rand index	0.98	0.97
Purity	97.5%	98.2%
NMI	0.96	0.98
CP	0.01	0.02
SP	0.97	0.97
DBI	0.02	0.03
DVI	0.96	0.98

TABLE II. COMPARISON OF RAND INDEX RESULTS OF DIFFERENT METHODS IN DIFFERENT DIMENSIONS

group	Methods in this paper		Literature [6] Method		Literature [7] Method	
	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set
1	0.98	0.97	0.89	0.87	0.88	0.81
2	0.99	0.96	0.88	0.88	0.81	0.75
3	0.97	0.98	0.85	0.84	0.86	0.78
4	0.98	0.97	0.86	0.85	0.72	0.79
5	0.99	0.97	0.89	0.82	0.83	0.82

IV. CONCLUSION

In order to deeply analyze the large amount of information generated by the simulation laboratory, it is necessary to accurately partition the information. An adaptive partition algorithm of simulation laboratory information set based on project-based learning data-driven and random grid is studied. The fuzzy C-means clustering algorithm in project-based learning data drive preliminarily divided the information, avoiding the constraints of time, space and other conditions. In the information space, the grid division algorithm of the data histogram was used to further divide the information, weakening the correlation between the data. Then, the grid division algorithm of the data histogram was used to cluster similar data, thus further enhancing the classification effect of the data. Through experiments, it is verified that the method can still maintain high classification accuracy in different dimensions, and has good application effect. However, there are still some shortcomings in this paper. Some heterogeneous data may appear in the data generated by the simulation laboratory, which is easy to be excluded as abnormal data when classifying. In future research, it is also necessary to strengthen the research on the classification and re clustering of heterogeneous data.

ACKNOWLEDGMENT

The study was supported by Shaanxi Province Education Science Planning Project --- “The research and practice of Project-based learning in higher education for English learning to improve the key competencies (Grant No. JYTYB2022-89)”, and Shaanxi Higher Education Association Project --- “Research on education and teaching reform of non-government undergraduate colleges under the background of new liberal arts” (Grant No.XGHZ2102).

REFERENCE

[1] P. Sidjanin, J. Plavsic, I. Arsenic, and M. Krmar, “Virtual reality (vr) simulation of a nuclear physics laboratory exercise,” European Journal of Physics, 2020, 41(6).

[2] S. He, D. Kong, J. Yang, L. Ma, and Y.Chang, “Research on the teaching mode of university virtual laboratory based on component technology,” International Journal of Continuing Engineering Education and Life-Long Learning, 2021, 31(1), 1.

[3] M. D. Koretsky, “An interactive virtual laboratory addressing student difficulty in differentiating between chemical reaction kinetics and equilibrium,” Computer applications in engineering education, 2020, 28(1), 105-116.

[4] L. F. Zapata-Rivera, and C. Aranzazu-Suescun, “Enhanced virtual laboratory experience for wireless networks planning learning,” Revista Iberoamericana de Tecnologias del Aprendizaje, 2020, PP(99), 1-1.

[5] M. Ricaurte, and A. Viloria, “Project-based learning as a strategy for multi-level training applied to undergraduate engineering students - sciencedirect,” Education for Chemical Engineers, 2020, 33, 102-111.

[6] A. Sutagundar, and P. Sangulagi, “Fog computing based information classification in sensor cloud- agent approach,” Expert Systems with Applications, 2021, 182(2), 115232.

[7] J. Flisar, and V. Podgorelec, “Improving short text classification using information from dbpedia ontology,” Fundamenta Informaticae, 2020, 172(3), 261-297.

[8] T. Zheng, Z. Duan, J. Wang, G. Lu, and Z. Yu, “Research on distance transform and neural network lidar information sampling classification-based semantic segmentation of 2d indoor room maps,” Sensors, 2021, 21(4), 1365.

[9] N. Berente, S. Seidel, and H. Safadi, “Data-driven computationally intensive theory development. Information Systems Research,” 2019, 30(1), 50-64.

[10] Y. L. Kang, L. L. Feng, and J. A. Zhang, “Cloud-Based Big Data Fuzzy Clustering Method Simulation Based on Grid Index,” Computer Simulation, 2019, 36(12):341-344+441.

[11] H. Liu, and Q. Qian, “Bi-level attention model with topic information for classification,” IEEE Access, 2021, PP(99), 1-1.

[12] Y. Song, L. Gao, X. Li, and W. Shen, “A novel point cloud encoding method based on local information for 3d classification and segmentation,” Sensors, 2020, 20(9), 2501.

[13] M. R. Bouadjenek, S. Sanner, and Y. Du, “Relevance- and interface-driven clustering for visual information retrieval,” Information Systems, 2020, 94(6), 101592.

[14] Z. Cai, X. Yang, T. Huang, and W. Zhu, “A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering,” Information Sciences, 2020, 508, 173-182.

[15] M. B. Hajkacem, C. B. N'Cir, and N. Essoussi, “One-pass mapreduce-based clustering method for mixed large scale data,” Journal of Intelligent Information Systems, 2019, 52(3), 619-636.

[16] F. Cicalese, and E. S. Laber, “Information theoretical clustering is hard to approximate,” IEEE Transactions on Information Theory, 2020, PP(99), 1-1.

[17] X. Zhang, W. Pan, Z. Wu, J. Chen, and R. Wu, “Robust image segmentation using fuzzy c-means clustering with spatial information based on total generalized variation,” IEEE Access, 2020, PP(99), 1-1.

[18] C. Wu, and X. Zhang, “A novel kernelized total bregman divergence-driven possibilistic fuzzy clustering with multiple information constraints for image segmentation,” IEEE Transactions on Fuzzy Systems, 2021, PP(99), 1-1.

[19] D. Wei, Z. Wang, L. Si, C. Tan, and X. Lu, “An image segmentation method based on a modified local-information weighted intuitionistic fuzzy c-means clustering and gold-panning algorithm,” Engineering Applications of Artificial Intelligence, 2021, 101(3), 104209.

[20] M. S. Talib, A. Hassan, T. Alameri, Z. A. Abas, and N. Ibrahim, “A center-based stable evolving clustering algorithm with grid partitioning and extended mobility features for vanets,” IEEE Access, 2020, PP(99), 1-1.