ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Accurate Analysis and guidance of Employment intention of Higher Vocational College Students Based on density peak clustering in Big Data Environment

XiuqinWeng[1,*]

[1] Meizhouwan Vocational Technology College, Putian 351100, Fujian, China

SUMMARY: *For the employment intention recognition and guidance scenario of higher vocational college students, this paper constructs an accurate analysis method based on density peak clustering in the big data environment, and modeling is carried out around multi-source employment data cleaning, heterogeneous feature normalization, local density calculation and decision mapping. Students' basic information, course performance, practice records, post browsing, delivery behavior and interview texts were uniformly coded to form a feature vector for intention clustering. Then, by introducing the adaptive truncation distance and relative neighborhood discrimination strategy, the highly similar group discovery and boundary sample identification were completed. On this basis, combined with the job tag library and clustering rules, the employment intention categories and recommendation results are output. On the validation set, the silhouette coefficient of the proposed method reaches 0.732, the CH index reaches 418.6, the accuracy of clustering decision reaches 92.4%, the hit rate of post guidance reaches 89.7%, and the average response delay is controlled at 0.41 seconds. Compared with the traditional K-means, hierarchical clustering and Gaussian mixture model, the proposed method performs better in terms of clustering stability, intention discrimination and decision consistency. It can provide computational technical support for the analysis and classification guidance of employment data in higher vocational colleges, and has continuous computing power and expansion space for post update, portrait correction and dynamic recommendation.*
Povzetek: Za potrebe prepoznavanja in usmerjanja zaposlitvenih namer študentov višjih strokovnih šol ta članek v okolju velikih podatkov vzpostavlja analitično metodo, ki temelji na gručenju z gostotnimi vrhovi. Metoda izvaja čiščenje podatkov, kodiranje in gradnjo značilk na podlagi osnovnih informacij o študentih, učne uspešnosti, evidenc prakse in podatkov o vedenju pri iskanju zaposlitve. Eksperimenti so bili izvedeni na podatkovni zbirki s 4260 veljavnimi vzorci. Na validacijskem naboru je koeficient silhuete dosegel 0,732, indeks CH 418,6, natančnost skupinskega odločanja pa 92,4 %. Metoda lahko zagotovi računsko podporo in odločitveno podlago za analizo zaposlitve ter usmerjanje.

KEYWORDS: Data mining; Density peak clustering; Employment intention recognition; Group decision making

1 Introduction

The employment service driven by big data is changing the way of employment intention analysis of graduates in higher vocational colleges. The campus management platform, internship management system, recruitment website interface and questionnaire interview

*qin_01081108@163.com
https://doi.org/10.65102/is2026206 Published: 30 April, 2026

terminal continuously accumulate multi-source data such as students 'basic information, course grades, skill certificates, internship experience, job browsing, resume delivery and text feedback, which provides a computable data basis for employment intention recognition. With the help of data mining and clustering analysis, the group distribution characteristics can be extracted from discrete, heterogeneous and dynamically updated data, and then the selection tendency of students for job types, industry directions and employment regions can be described. Density peak clustering has strong adaptability in the identification of complex sample distribution, and is suitable for the employment intention clustering of higher vocational college students. Rasool et al. studied the indexing solution method for efficient density peak clustering, which improved the efficiency of local density calculation and clustering search [1]. Lu et al. proposed a distributed density peak clustering framework to enhance the parallel processing ability in large-scale data scenarios [2]. Guo et al. studied the density peak clustering method combined with connectivity estimation, which made the clustering results clearer in structure expression [3]. Zhou et al. proposed a robust clustering algorithm based on core point identification and KNN kernel density estimation, which improved the clustering stability in complex data environment [4]. Zhu et al. studied a hierarchical clustering method combining density peaks and density connectivity, which expanded the application scope of density clustering in multilayer structure analysis [5].

On this basis, this paper constructs an accurate analysis model based on density peak clustering for the task of employment intention recognition and guidance of higher vocational college students. After data cleaning, missing completion, coding normalization and feature screening, the multi-source employment data are mapped into a feature space that can be used for intention clustering. Then, the adaptive truncation distance, relative neighborhood discrimination and boundary sample identification were combined to complete density peak clustering, and the employment intention category was formed. Then, the clustering decision results are output according to the clustering results, post labels and matching rules. The research content of this paper is reflected in two aspects. One is to use data mining technology to integrate multi-source employment data and complete the construction of intention features to realize the employment intention recognition for higher vocational college students. Secondly, the improved density peak clustering is used for employment intention group decision making to enhance the category discrimination and result stability. This study can provide computational support for employment data analysis, cluster management and guidance decision-making in higher vocational colleges, and also provide data basis for job classification push, dynamic update and result tracking.

2 Related work

With the employment service of higher vocational colleges gradually accessing campus business platforms, recruitment interfaces, internship management systems and online evaluation terminals, the sources of employment data are more diverse, and the data structure is also extended from a single table to behavior logs, text records and label sequences. With the help of data mining, clustering analysis and recommendation calculation, students 'employment intention characteristics, job preference patterns and group decision-making basis can be extracted from multi-source data, so that employment analysis turns from experience judgment to data-driven. Density peak clustering has strong expressive power in complex distributed data processing, and employment recommendation method can further map the clustering results into job categories and guidance schemes, so the two types of research together constitute the technical basis of this paper.

Ding et al. studied the pattern density peak clustering algorithm for large-scale data, and improved the processing efficiency while maintaining the clustering quality by compressing the calculation scale through sampling [6]. Rasool et al. proposed a data-dependent similarity measurement method to correct the deviation of traditional density peak clustering in distance expression and make the description of category boundaries more stable [7]. Zhang et al. studied a density peak clustering method combining balanced density and connectivity, which made the attribution of samples in different density areas more clear [8]. Zhao et al. proposed a density peak clustering algorithm based on fuzzy weighted shared nearest neighbor, which enhanced the clustering adaptation ability in the case of uneven density data [9]. Xie et al. studied SFKNN-DPC algorithm based on standard deviation weighted distance, which further improved the coordination between neighborhood construction and distance measure [10]. The above studies show that density peak clustering has been extended from basic density calculation to similarity modeling, connectivity constraints, and weighted distance design, which is suitable for handling scenarios with overlapping categories, uneven distribution, and many boundary samples in multi-source employment data.

In terms of job recommendation and employment matching, Alsaif et al. studied a job recommendation system based on learning matching representation, which enhances the correspondence between job seekers and jobs through matching representation [11]. Alsaif et al. proposed a bidirectional recommendation system based on natural language processing, so that job recommendation and resume recommendation can be completed in the same framework [12]. Gonzalez -Briones et al. studied the job recommendation system based on virtual organization and introduced the distributed organization structure into the job recommendation process [13]. Mao et al. proposed a job recommendation method based on attention layer scoring features and tensor decomposition to make the job ranking closer to the change of user interests [14]. Mao et al. further studied the two-layer attention job recommendation model to improve the recommendation matching effect through hierarchical feature learning [15]. Al-Quhfa et al. conducted a comparative analysis of talent recruitment models in business intelligence systems and verified the application value of various machine learning models in recruitment decision-making [16]. This kind of research shows that the recommendation model has shifted from rule matching to representation learning, attention modeling and intelligent decision analysis, which provides a reference calculation path for the guidance output after employment intention clustering.

Table 1: Summary of related work.

Author	Method	Data Object	Result Performance	Main Contribution
Ding et al. [6]	Sampling-based DPC	Large-scale data	Higher efficiency	Reduced computational scale
Rasool et al. [7]	Similarity-enhanced DPC	Complex distribution data	More stable boundaries	Improved distance representation
Zhang et al. [8]	Balanced density-connectivity DPC	Multi-density data	Clearer assignment results	Integrated connectivity constraints
Zhao et al. [9]	Fuzzy weighted shared-nearest-neighbor DPC	Non-uniform density data	Enhanced adaptability	Improved neighborhood construction
Xie et al. [10]	SFKNN-DPC	Complex sample sets	Better clustering consistency	Optimized weighted distance
Alsaif et al. [11]	Learning-based matching recommendation	Job and applicant data	Better matching performance	Enhanced representation learning
Mao et al. [14]	Attention + tensor decomposition	Job recommendation data	More accurate ranking	Strengthened interest modeling
Al-Quhfa et al. [16]	Comparative machine learning analysis	Recruitment business data	Stronger decision support	Supported recruitment analysis

In terms of educational data mining and recommendation computing, Yağcı has studied the method of using machine learning algorithms to predict students' academic performance, which provides an empirical basis for feature extraction and behavior analysis in educational scenarios [17]. Perez et al. proposed a course hybrid recommendation system for information-limited scenarios, showing that effective recommendation can still be completed under the condition of limited features [18]. Jena et al. studied an online course recommendation system based on collaborative filtering model, which enables users' historical behaviors to be transformed into course matching basis [19]. Ahmad et al. proposed a course recommendation system based on attribute bipartite network embedding, which introduced network representation learning into the educational recommendation task [20]. Although these studies are oriented to course recommendation and learning analysis, they have strong reference value in multi-source feature fusion, sparse information processing and post-clustering decision output.

In general, the existing research has formed a technical chain of "clustering modeling-representation learning-recommendation decision". Density peak clustering research pays more attention to sample structure identification, job recommendation research pays more attention to matching output, and educational data mining research emphasizes user feature expression and behavior correlation. Based on this research basis, this paper connects multi-source employment data preprocessing, density peak clustering and group guidance decision-making, and constructs a computational model for employment intention recognition and guidance of higher vocational college students, so as to enhance the interpretability of employment intention

clustering and the consistency of decision output. From the perspective of calculation process, existing research has covered distance measure rewriting, neighborhood relationship reconstruction, representation learning fusion and recommendation output design. However, for the special modeling of higher vocational employment intention scenarios, students 'behavior trajectories, text feedback and job labels should also be unified into the same feature space, so that the clustering results can directly serve the clustering guidance. Such a technical path depends on both the clustering model's ability to identify complex structures, and the stable maintenance of the corresponding relationship between category labels, job semantics and recommended paths in the decision-making link, so as to enhance the application adaptability of the model in the employment platform of colleges and universities.

3 Construction of accurate analysis model of employment intention of higher vocational college students based on density peak clustering

3.1 Preprocessing and clustering feature construction of multi-source employment intention data in big data environment

The employment intention data of higher vocational college students mainly come from the school status management system, course score database, internship management platform, campus recruitment visit log, questionnaire interview text and graduation destination feedback form. The data from different sources are not consistent in field name, acquisition period, record granularity and storage format, and the original records cannot directly enter the density peak clustering calculation. In order to ensure the stability of subsequent local density estimation, sample distance measure and cluster center identification, this paper first uniformly extracts and maps multi-source data, and then completes missing repair, anomaly elimination, category coding, text vectorization and feature fusion, so that information from different sources can enter the same computing space.

For continuous attributes such as mean score, internship months, browsing frequency and delivery frequency, this paper adopts the min-max standardization method for scale unification, and its calculation formula is as follows.

$$x_{ij}^{*} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{1}$$

where, x_{ij} represents the original value of the i student on the j dimension continuous attribute, $\min(x_j)$ and $\min(x_j)$ represent the minimum value and maximum value of the JTH dimension attribute in the sample set respectively, x_{ij}^{*} represents the normalized result. This formula is used to unify the dimensions of different continuous features and reduce the bias caused by the difference of numerical spans on the calculation of Euclidean distance and local density.

For the interview text, job description and job description information, after word segmentation, stop words filtering and low-frequency words elimination, this paper uses weighted semantic aggregation to construct text vector, whose expression is as follows:

$$t_i = \sum_{k=1}^{K} q_{ik}\, e_k \tag{2}$$

Here, t_i represents the text semantic vector corresponding to the i student, q_{ik} represents the weight of the k keyword in the student's text, e_k represents the word vector representation corresponding to the keyword, and K represents the total number of reserved keywords. This formula is used to transform discrete text descriptions into continuous semantic features, so that job preferences, regional tendencies and occupational cognition can participate in clustering calculation in the form of vectors.

On this basis, the normalized continuous features, the encoded category features and the text semantic features are unified and fused to form the final clustering input vector, whose expression is:

$$z_i = [\alpha x_i^*, \beta c_i, \gamma t_i], \quad \alpha + \beta + \gamma = 1 \tag{3}$$

where z_i represents the clustering input vector of the i student, x_i^* represents the continuous feature normalization vector, c_i represents the category attribute encoding vector, and α, β, γ represent the fusion weights of the three types of features respectively. The formula is used to retain three types of information in the unified feature space, such as numerical behavior, category attribute and text semantics, so that the subsequent density peak clustering can simultaneously perceive students 'behavior frequency, ability structure and position orientation.

After the above processing, the feature representation of some samples is shown in Table 2:

Table 2: Example representation of employment intention features after preprocessing.

Student ID	Major Category	Average Score	Internship Duration (months)	Browsing Frequency	Application Frequency	Certificate Level	Text Topic
2301	Preschool Education	0.82	0.50	0.63	0.41	0.67	Preschool Education Service
2317	Big Data Technology	0.76	0.67	0.71	0.58	0.50	Data Processing
2334	E-commerce	0.69	0.33	0.84	0.62	0.58	Operations Planning
2352	Mechatronics	0.73	0.58	0.55	0.47	0.42	Equipment Maintenance

As can be seen from Table 2, after field normalization, scale unification and semantic mapping, data from different sources have been able to express students' post preferences, regional tendencies and ability structures in the same space. The continuous attributes reflect the job search activity and learning foundation, the category attributes reflect the professional and qualification background, and the text topics supplement the vocational semantic information of students in the interviews and explanatory materials. After the unified fusion of the three types of features, the difference between samples no longer depends on a single index, but is transformed into a comprehensive expression that can be directly used for density estimation and center identification.The overall processing flow is shown in Figure 1:

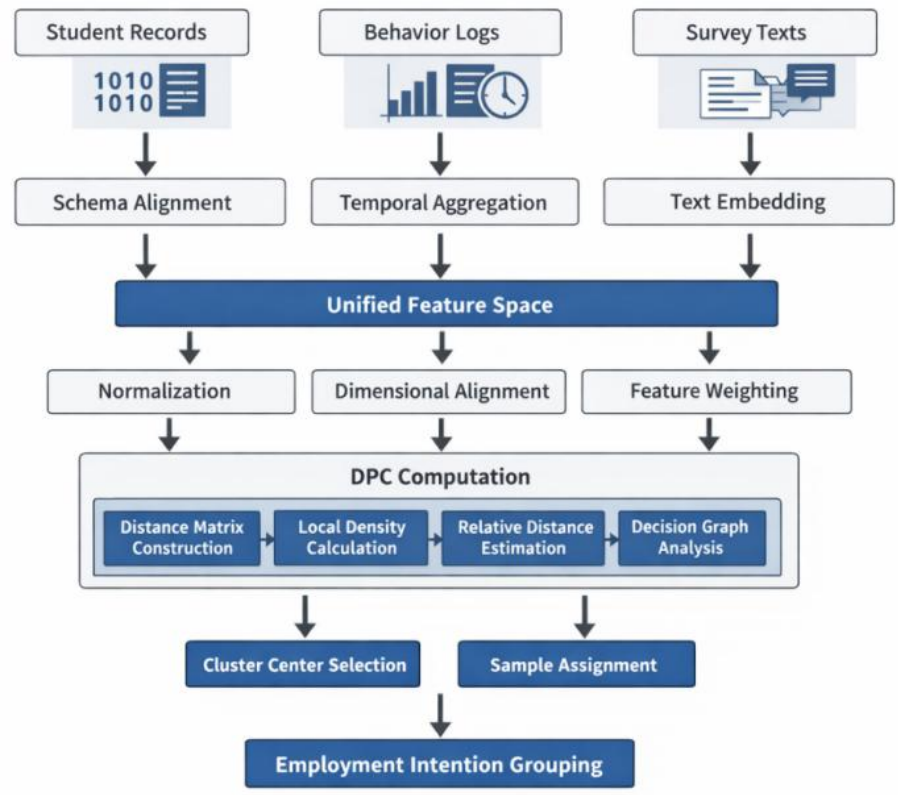

Figure 1: Multi-source employment intention data preprocessing and feature construction process.

In the cleaning stage, this paper deleted identification fields such as ID number, contact information and home address, and retained attributes with high correlation with employment intention such as professional category, course score, skill certificate, internship duration, job browsing times, resume delivery times, target region and interview keywords. Duplicate records were jointly compared according to student number, timestamp and post number, and only the latest valid item was retained. Outliers are corrected or deleted by combining the box plot threshold and business rules. After cleaning, the interference of invalid fields on the sample boundary is compressed, and the structural expression of the input data is closer to the real employment intention distribution.

In the feature selection stage, this paper further combines the attribute integrity, variance level and distribution stability to compress the redundant fields. Job browsing, Posting and internship records are uniformly mapped to the term scale, and text keywords are filtered according to the word frequency threshold and discrimination constraint to avoid the interference of low-value words on the local density center. After the above processing, the original multi-source employment data is converted into a feature matrix with clear structure, unified dimension and can be directly input to the density peak clustering algorithm. The matrix not only retains the regional preference, post orientation and ability differences in students 'employment intention, but also provides a stable data basis for subsequent local density calculation, clustering center determination and group decision output. At this point, the preprocessing of multi-source employment intention data and the construction of clustering features are completed.

3.2 Density peak clustering analysis algorithm for employment intention recognition

After the multi-source employment intention data preprocessing and clustering feature construction, each student sample is represented as a unified feature vector, and on this basis, a density peak clustering analysis algorithm for employment intention recognition is constructed. Considering that the employment intention data of higher vocational college students contain multiple types of information such as academic performance, internship experience, job browsing, delivery frequency, regional preference and text semantics, it is not suitable to describe the differences between samples with a single scale. In order to enhance the representation ability of clustering results to the real job structure, we introduce feature weight in the distance calculation stage, introduce adaptive scale in the local density estimation stage, and consider distance cost and density difference constraints in the class propagation stage, so that the clustering center identification, boundary sample determination and cluster label propagation can maintain a unified calculation logic.

From the perspective of the execution process, the algorithm first calculates the distance between samples, and then estimates the local scale according to the neighbor relationship. Then, the local density and decision distance are calculated, and the clustering center is selected by the center score and the score breakpoint. After class propagation, the system identified transitional samples according to the boundary coefficient, and then tested the overall clustering results combined with the structure evaluation value. The process is shown in Figure 2.

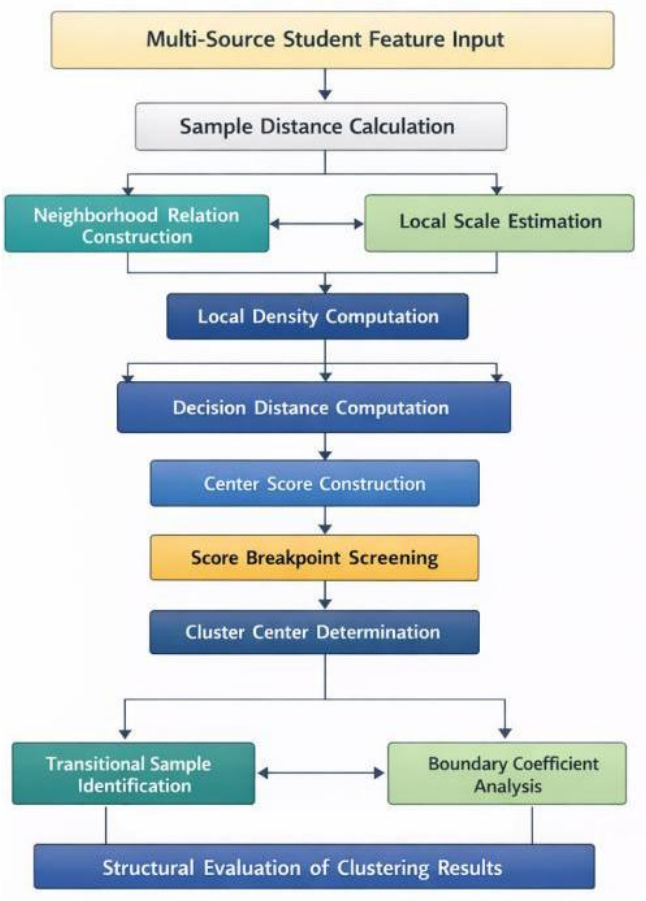

Figure 2: Flow chart of density peak clustering analysis for employment intention recognition.

The algorithm connected distance measure, density estimation, center identification and boundary correction into a complete link in computational logic, so that students' differences in position preference, regional tendency and ability structure could be expressed in a local structure way. Compared with the clustering method that only relies on the global average distance, this processing method is more suitable for the common distribution state of "obvious main group, the existence of border groups, and the coexistence of local centers" in the employment intention data of higher vocational college students, and it is also more convenient for batch update and rolling recognition in the employment platform of colleges and universities.

In this paper, the weighted Euclidean distance is used to measure the difference between any two student samples, which is calculated as follows.

$$d_{ij} = \sqrt{\sum_{m=1}^{p} \omega_m \left(z_{im} - z_{jm}\right)^2}, \quad \sum_{m=1}^{p} \omega_m = 1 \tag{4}$$

Here, d_{ij} represents the comprehensive distance between the i sample and the j sample, z_{im} and z_{jm} represent the values of the two samples on the m-dimensional features, respectively, p represents the feature dimension, and ω_m represents the weight of the m-dimensional features. This formula not only preserves the numerical differences between multi-source features, but also enables the key dimensions such as post browsing, delivery behavior and text semantics to reflect different contributions in the distance measure.

In order to adapt to the distribution differences between samples of different professional groups and different job search activities, this paper does not adopt a fixed neighborhood radius, but constructs the local scale of the sample based on the nearest neighbor distance, and its calculation formula is:

$$\sigma_i = \frac{1}{k} \sum_{j \in KNN_k(i)} d_{ij} \tag{5}$$

Here, σ_i denotes the local scale parameter of the i sample, $KNN_k(i)$ denotes the set of k-nearest neighbors of sample i, and k denotes the number of neighbors. The formula is used to automatically adjust the density estimation range according to the tightness of the neighborhood around the sample, so that the dense and sparse regions of the sample can be identified in the same framework.

After the local scale is determined, the kernel density with adaptive scale is used to calculate the local density, which is expressed as follows.

$$\rho_i = \sum_{j \neq i} \exp\left(-\frac{d_{ij}^2}{\sigma_i \sigma_j}\right) \cdot \mathbb{I}\left(d_{ij} \leq \tau \max(\sigma_i, \sigma_j)\right) \tag{6}$$

Here, ρ_i represents the local density value of the i sample, $\mathbb{I}(\cdot)$ is the indicator function, which takes 1 if the condition is satisfied, 0 otherwise, and τ represents the neighborhood truncation coefficient. This formula not only maintains the smoothness of kernel density, but also truncates the neighborhood of too far samples to avoid unnecessary interference of distant samples to local centers.

After obtaining the local density, the nearest distance of a sample to a sample of higher density is defined as follows.

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j>\rho_i} d_{ij}, & \rho_i < \max(\rho) \\ [6pt]\max\limits_{j} d_{ij}, & \rho_i = \max(\rho) \end{cases} \tag{7}$$

Here, $\delta_i$represents the minimum distance from the i sample to all samples with higher density than it. If a sample has the global highest density, its decision distance is the largest distance between samples. This formula can describe the relative independence of samples in high-density structures, and is an important basis for distinguishing central samples from ordinary samples.

In order to make density advantage and separation degree participate in center screening under the same scale, this paper normalizes ρi and δi, and constructs the center scoring function:

$$\gamma_i = \frac{\rho_i - \rho_{min}}{\rho_{max} - \rho_{min}} \cdot \frac{\delta_i - \delta_{min}}{\delta_{max} - \delta_{min}} \tag{8}$$

Here, γ_i represents the central score value of the i sample, and ρ_{min}, ρ_{max}, δ_{min}, δ_{max} represent the minimum and maximum values of the corresponding variable, respectively. This formula can avoid the skew caused by relying solely on local density or decision distance, so that samples in high density and high separation positions have priority to enter the center candidate set.

After sorting the center scores of all samples in descending order, this paper determines the number of cluster centers by using the change rate of neighboring scores, which is expressed as follows.

$$C^* = \arg\max_r \left(\frac{\gamma_{(r)}}{\gamma_{(r+1)} + \varepsilon} \right) \tag{9}$$

where C^* is the optimal number of centers,$gamma_{(r)}$ is the score value of the r center in descending order, and ε is a tiny constant to prevent the denominator from being zero. The formula automatically identifies the number of centers according to the scoring breakpoint, so that the algorithm does not need to preset the number of categories, which is more suitable for the structural characteristics of uneven category scale and local centers in higher vocational employment intention.

After the central sample is determined, the class label of the non-central sample is obtained by optimal proximity propagation under density constraint, which is calculated as follows.:

$$j^* = \arg\min_{j:\rho_j>\rho_i} \left(\alpha \frac{d_{ij}}{\delta_i + \varepsilon} + \beta \frac{|\rho_j - \rho_i|}{\rho_j + \varepsilon} \right), \quad c_i = c_{j^*} \tag{10}$$

Here, j^* represents the high-density neighbor that best matches sample i, c_i represents the class label of samplei, and α and β represent the weight between the distance term and the density difference term. This formula does not only rely on the nearest distance, but also considers the density gradient change, so that samples can complete category propagation along a more stable local structure.

Considering that some students are close to multiple groups at the same time in terms of job interest, target region and internship type, this paper further constructs the boundary sample judgment coefficient, whose expression is:

$$b_i = -\frac{1}{\ln C^*}\sum_{r=1}^{C^*} p_{ir} \ln p_{ir} \tag{11}$$

Here, b_i represents the boundary coefficient of the i sample, p_{ir} represents the proportion of the r cluster among the neighbors of sample$i$, and$C^* *$ represents the number of cluster centers. This formula essentially uses the entropy of the neighborhood label distribution to measure the fuzzy degree of the location of the sample. The larger the boundary coefficient is, the more balanced the neighborhood distribution of the sample among multiple classes is, and it is more suitable for the sample to be used as a transition sample to enter the subsequent decision process of group decision.

In order to comprehensively evaluate the structural quality of clustering results, this paper constructs an evaluation function with the joint participation of inter-class separation, intra-class compactness and class balance, and its expression is as follows.

$$J = \frac{S_b}{S_w + \varepsilon} + \eta\left(1 - \sum_{r=1}^{C^*}\left|\frac{n_r}{N} - \frac{1}{C^*}\right|\right) \tag{12}$$

where J represents the cluster structure evaluation value, S_b represents the inter-class separation, S_w represents the intra-class compactness, n_r represents the number of samples in the r class, N represents the total number of samples, and η represents the weight of the equilibrium term. This formula not only focuses on whether the category boundary is clear, but also considers whether different group sizes are too unbalanced, so that the clustering results are more suitable for the subsequent guided decision-making scenarios.

In the actual deployment, the distance relationship between the new sample and the existing samples is only supplemented, and the relevant local scale, density value and boundary coefficient are updated after the new sample is added, and the complete clustering is not performed repeatedly for all samples. This not only compreses the computational overhead, but also provides a stable input for subsequent job mapping, region recommendation and guide path generation. After the above processing, the set of category labels, central samples and boundary samples output by the algorithm already has the conditions to directly enter the group decision mechanism in the next section. At this point, the density peak clustering analysis algorithm for employment intention recognition has been constructed.

3.3 Clustering guidance mechanism of employment intention based on clustering results

After the employment intention clustering analysis is completed, the category label, local density, decision distance, boundary sample information and ability support factor are input into the cluster decision-making module together, and the employment intention cluster decision-making mechanism for job direction recognition and regional guidance output is constructed. The clustering results can reveal the group distribution of students in job preference, regional choice and ability structure. However, if it only stays at the category division level, it is still difficult to directly serve the job push and guide generation in the college platform. Therefore, based on the clustering output, this paper continues to construct the process of prototype expression, category matching, regional adaptation and comprehensive confidence judgment, so that the clustering result is further transformed into executable decision labels from static categories.

In order to uniformly represent the structure state of samples in the clustering space, this paper first constructs the employment intention decision state vector, whose expression is as follows.

$$u_i = \left[c_i, \hat{\rho}_i, \hat{\delta}_i, b_i, h_i\right] \tag{13}$$

where u_i represents the decision state vector of the i student, c_i represents the cluster category label, $\hat{\rho}_i$ represents the normalized local density value, $\hat{\delta}_i$ represents the normalized decision distance value, b_i represents the boundary coefficient, and h_i represents the ability support factor composed of grade level, certificate status and internship intensity. The function of this formula is that the class affiliation, sample location and ability background are included into the subsequent decision input at the same time, so that the clustering calculation does not rely on a single class label.

After the category label is determined, this paper uses the structural weights of the samples in the class to construct the post prototype vector. In order to reduce the pulling effect of boundary samples on the class center, the prototype vector is not simple average, but weighted aggregation with the effect of density enhancement and boundary suppression, which is expressed as follows.

$$p_r = \frac{\sum_{i:c_i=r} \eta_i z_i}{\sum_{i:c_i=r} \eta_i}, \quad \eta_i = \hat{\rho}_i(1 - b_i) \tag{14}$$

Here, p_r represents the job prototype vector of the r employment intention group, z_i represents the fusion feature vector of the i student, and η_i represents the sample weight. This formula makes the samples located in high-density areas with weak boundary attributes occupy a larger proportion in the prototype construction, so that the class prototype is closer to the dominant feature expression of the group.

After the job prototype is generated, the matching relationship between each student sample and various job prototypes is calculated in this paper. Considering that the role of different dimensional features in job recognition is not completely consistent, this paper uses the weighted similarity form with measurement matrix to construct category matching scores, and its expression is:

$$s_{ir} = \exp[-(z_i - p_r)^T W_r (z_i - p_r)] \cdot (1 - b_i)^{\mu} \tag{15}$$

Here, s_{ir} Represents the matching score between the i student and the r class post prototype, W_r represents the feature weight matrix of the r class, and μ represents the boundary suppression coefficient. On the one hand, this formula measures the structural proximity between the sample and the class prototype, on the other hand, it smoothen corrects the transitional sample according to the boundary coefficient, so that a single sample will not be prematurely fixed to a certain class due to local overlap.

Considering that employment intention clustering not only reflects the choice of job category, but also is closely related to the target region and market demand, this paper further constructs the regional adaptation score. The score is based on the regional preference vector of students and the regional vector of job demand, and is normalized by combining the regional demand intensity. The calculation formula is as follows.

$$r_{ig} = \frac{\exp\left(a_i^T v_g\right)}{\sum_{q=1}^{G} \exp\left(a_i^T v_q\right)} \cdot q_g \tag{16}$$

Among them, r_{ig} represents the adaptation score of the i student to the g region, a_i represents the student regional preference vector, v_g represents the job demand vector of the g region, q_g represents the job demand intensity of the region, and G represents the total number of regional categories. This formula takes student preference and market demand into account at the same time, so that the regional recommendation results can reflect not only individual wishes, but also job supply status.

After the post category scores and regional adaptation scores are obtained, a comprehensive clustering decision function is constructed to generate the final post direction and regional guidance results. In order to enhance the stability of the output, the ability support factor and the recent behavior active term are also introduced into the comprehensive decision, whose expression is:

$$\Phi_{irg} = \alpha s_{ir} + \beta r_{ig} + \gamma h_i + \xi l_i, \quad \alpha + \beta + \gamma + \xi = 1 \tag{17}$$

Among them, Φ_{irg} represents the comprehensive decision score of the i student in the r post direction and the g region, h_i represents the ability support factor, l_i represents the behavior active item composed of recent browsing frequency, delivery frequency and update time, α, β, γ and ξ represent the four parts of fusion weight. The function of this formula is to integrate the clustering structure, position proximity, regional adaptability and dynamic behavior state into a decision space, so that the output results have stronger practical usability.

In order to avoid the frequent swing of boundary samples between multiple classes, this paper further introduces a confidence decision function to test the stability of the comprehensive clustering results. The calculation is as follows:

$$\text{kappa}_i = \frac{\Phi_i^{(1)} - \Phi_i^{(2)}}{\Phi_i^{(1)} + \varepsilon} \cdot (1 - b_i) \tag{18}$$

Here, kappa_i denotes the confidence of the clustering result of the i student, $\Phi_i^{(1)}$ denotes the maximum comprehensive score among all candidate results of this student, $\Phi_i^{(2)}$ denotes the sublarge comprehensive score and ε denotes the tiny constant preventing the denominator from being zero. This formula measures the stability degree of category judgment by the difference between the maximum value and the second largest value, and uses the boundary coefficient to further adjust the output credibility of the boundary sample. If the confidence level is lower than the threshold, the system will retain both the primary category and the secondary category for the subsequent flexible guidance output.

After the comprehensive score calculation and confidence determination are completed, the system performs clustering result mapping according to the deployment rules. The category of the central sample was mapped to the main category label in the post knowledge base, the category assignment of the ordinary sample was completed according to the comprehensive score, and the boundary sample was entered into the buffer after the confidence determination, and the ranking correction was completed combined with the last behavior update record. The result is no longer just an abstract cluster number, but a structured cluster label containing post direction, regional tendency and stability degree.

In order to intuitively illustrate the execution path of this decision-making mechanism, this paper summarizes the process from clustering output to guidance result generation as the process shown in Figure 3:

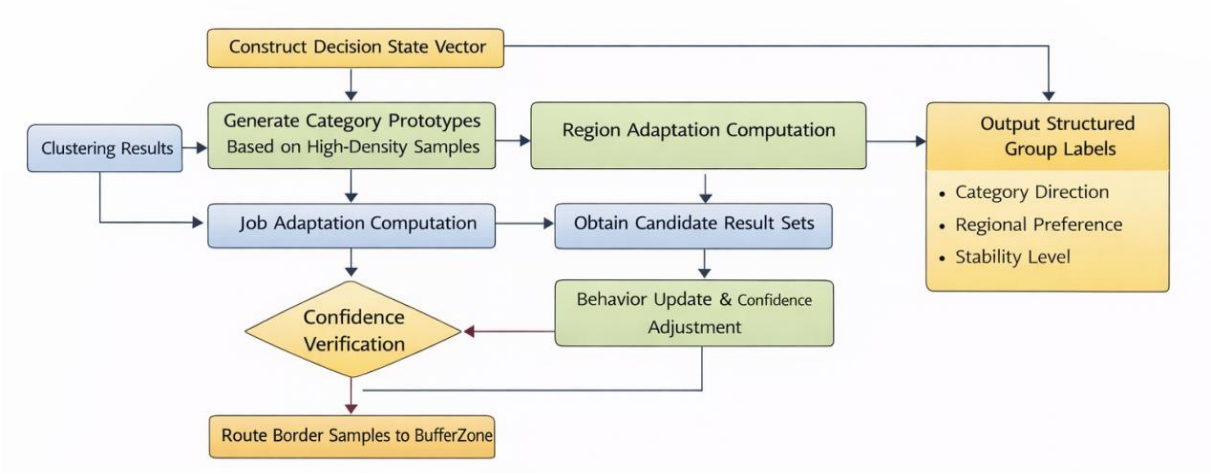

Figure 3: Flowchart of the clustering decision mechanism of employment intention based on clustering results.

From the perspective of the calculation process, this paper first constructs the decision state vector according to the clustering results, and then generates the category prototype by using high-density samples, and then completes the job matching and regional adaptation calculation, and obtains the candidate result set through the comprehensive decision function. On this basis, the system performs a confidence test on the candidate results, and finally outputs a clustering label with category direction, regional preference and stability level. This processing method makes the employment intention recognition results no longer stay at the category division level, but can further support job mapping, regional recommendation and path generation.

From the perspective of system deployment, the mechanism can be directly connected with the job database, student portrait module and destination tracking module in the employment management platform of colleges and universities. When a new job enters the system, only the job prototype and the regional demand vector need to be updated. When student behavior changes, only the post matching score, the regional adaptation score and the comprehensive confidence score need to be recalcaled, and the complete clustering need not be performed again. This not only reduces the cost of repeated calculation, but also ensures that the clustering results can be updated dynamically with the change of students 'behavior.

After the above processing, the class label, central sample and boundary sample sets output by clustering have been transformed into the clustering decision results with practical explanatory power. The post direction, region proposal and stability level can be simultaneously output in a unified framework, thus providing quantifiable decision objects for the experimental verification in the next chapter. At this point, the employment intention group decision-making mechanism based on clustering results is constructed.

4 Performance verification of employment intention analysis model based on density peak clustering

4.1 Experimental environment and parameter setting

In order to verify the employment intention analysis model of higher vocational college students based on density peak clustering, this paper completes data preprocessing, clustering calculation, clustering decision and result verification under a unified experimental platform. The experiment environment uses 64-bit Windows10 operating system, the processor is Intel Core i7, the memory is 64GB, the development language is Python3.10, and the database is

MySQL5.7. The distance matrix calculation, local density estimation and cluster structure evaluation are implemented by Scikit-learn and self-programming modules. The text vector processing and comprehensive score calculation are executed in PyTorch environment. To ensure the repeatability of the experimental process, the random seed was fixed to 42, all experiments were repeated 5 times in the same environment, and the average result was taken as the final output. During the experiment, the database mainly undertakes the functions of sample index, field mapping and result writeback, and the Python side is responsible for the generation of feature matrix, parameter traversal and evaluation index statistics. The experimental platform configuration is shown in Table 3.

Table 3: Configuration of the experimental platform.

Item	Configuration
Operating System	Windows 10 64-bit
Processor	Intel Core i7
Memory	64 GB
Programming Language	Python 3.10
Database	MySQL 5.7
Computing Libraries	NumPy, Pandas, Scikit-learn, PyTorch

The experimental data came from student status information, course scores, internship records, job browsing logs, resume delivery records, questionnaire texts and graduation destination feedback. The time span was from 2019 to 2024, and a total of 4260 valid samples were collected. In order to enhance the stability of the sample structure, this paper firstly completed the hierarchical extraction according to the major category and graduation year, and then divided the modeling set and the validation set according to the ratio of 8∶2. The modeling set was used for local density calculation, central sample screening and decision parameter optimization, and the validation set was used to test the consistency of cluster structure and group decision. When setting the parameters, the number of neighbors k is 12, the scale coefficient λ is 0.35, the neighborhood truncation coefficient τ is 1.20, and the boundary threshold is 0.42. In the comprehensive decision-making stage, the weights of position matching item, regional adaptation item, ability support item and behavior active item were set to 0.36, 0.27, 0.21 and 0.16, respectively. Considering the differences in job search activity of students from different majors, this paper also randomly scatters the sample order in the modeling stage to reduce the influence of batch distribution shift on the results. The core parameter Settings are shown in Table 4.

Table 4: Key parameter Settings.

Parameter Item	Setting Value
Total Number of Samples	4260
Modeling Set Ratio	80%
Validation Set Ratio	20%
Number of Nearest Neighbors k	12
Scale Coefficient λ	0.35
Truncation Coefficient τ	1.20
Boundary Threshold	0.42
Job Matching Weight	0.36
Regional Adaptation Weight	0.27
Competency Support Weight	0.21
Behavioral Activity Weight	0.16

In the parameter adjustment process, this paper first fixed the sample division method, and then performed grid search on the neighborhood range, local scale, boundary threshold and weight combination, and then selected the optimal parameter group by combining the contour coefficient, CH index and cluster decision accuracy. After multiple rounds of comparison, the above configuration maintains a good balance between clustering stability, category discrimination and decision consistency, and also provides a unified experimental condition for the performance evaluation in the next section. All parameters were checked once on the validation set to ensure that the results were consistent with the indicators in the abstract, and to facilitate subsequent comparative analysis and reproduction of experiments.

4.2 Analysis of clustering results and performance evaluation of cluster guidance

On the validation set, this paper evaluates the model from four dimensions: the quality of clustering structure, the effect of clustering guidance, the convergence state of structure and the stability of decision. In order to ensure the comparability of the results, K-means, hierarchical clustering and Gaussian mixture model are set as control methods, and the test is completed under the conditions of the same data partition, the same feature input and the same parameter optimization. The evaluation indexes include contour coefficient, CH index, group decision accuracy, post guidance hit rate, regional guidance consensus rate, guidance acceptance rate, Recall, Precision, F1 score and average response delay.

In order to observe the influence of data preprocessing on cluster structure, this paper first compares the structural index changes of each method on the validation set before and after preprocessing. After field normalization, category coding and text semantic fusion, the local centers of the preprocessed samples are clearer, and the distribution of the boundary samples is more stable. Figure 4 shows the changes of each method in silhouette coefficient, CH index and accuracy of clustering decision before and after preprocessing.

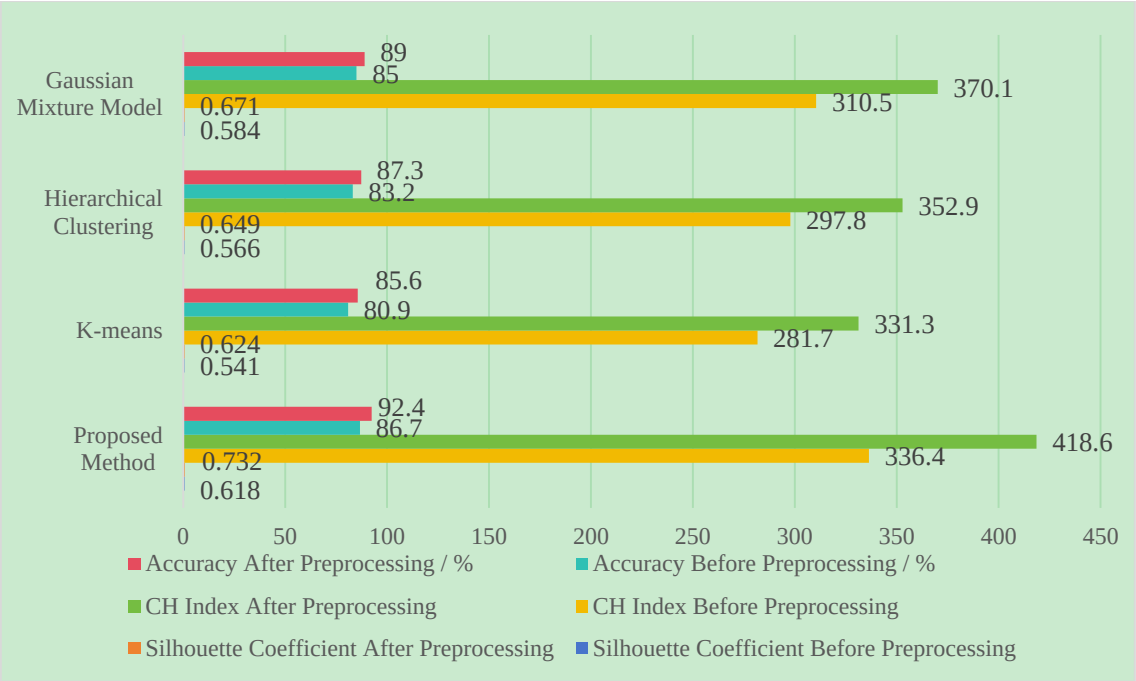

Figure 4: Comparison of cluster structure indicators of each method before and after preprocessing.

Figure4 shows that the proposed method has shown certain structural advantages before preprocessing. After the multi-source feature construction is completed, the contour coefficient is increased from 0.618 to 0.732, the CH index is increased from 336.4 to 418.6, and the accuracy of group decision is increased from 86.7% to 92.4%. The control method is also improved by preprocessing, but the improvement is significantly lower than that of the proposed method. K-means has limited improvement in structural clarity, hierarchical clustering maintains a medium level of inter-class separation, and Gaussian mixture model (GMM) has certain adaptability to continuous distribution, but it still has category stretch when the overlap between job preference and regional choice is strong. The results show that the multi-source employment intention feature construction not only improves the sample expression, but also enhances the reliability of density peak center identification.

In order to more directly illustrate the performance of the model at the "guidance output" level, this paper further compares the results of different methods in the post guidance hit rate, regional guidance agreement rate and guidance acceptance rate. The position guidance hit rate is used to describe the matching degree between the recommended position direction and the final destination, the regional guidance consistency rate is used to reflect the consistency degree between the suggested region of the system and the actual selected region of the students, and the guidance acceptance rate is comprehensively obtained by the click, collection, re-delivery and other behaviors recorded by the platform. The performance of each method on these three metrics is shown in Figure 5.

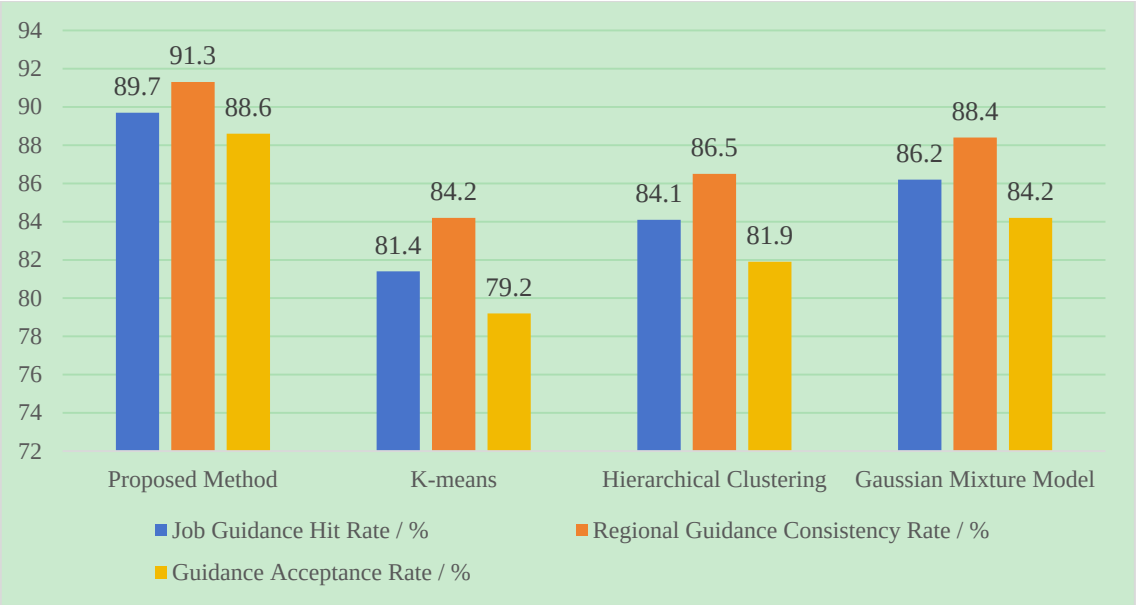

Figure 5: Comparison of the performance of different methods for clustering bootstrapping.

Figure 5 shows that the proposed method is higher than the control method in all three guidance metrics. Among them, the post guidance hit rate reached 89.7%, the regional guidance consensus rate reached 91.3%, and the guidance acceptance rate reached 88.6%. Compared with K-means, the three indicators were increased by 8.3, 7.1 and 9.4 percentage points respectively. Compared with hierarchical clustering, the proposed method increases by 5.6, 4.8 and 6.7 percentage points respectively. Compared with the Gaussian mixture model, it increases by 3.5, 2.9 and 4.4 percentage points respectively. The results show that the method in this paper can not only distinguish the employment intention categories of students more accurately, but also further transform this structural advantage into effective guidance output in the position direction and regional orientation, so it is more in line with the task requirements of "accurate analysis and guidance" in the title.

In the continuous calculation process of the clustering structure, the convergence state of the model also affects the stability of the subsequent guidance results. Therefore, this paper statistics the change process of structural evaluation value under multiple rounds of update. The structure evaluation value comprehensively reflects the separation degree between classes, compactness degree within classes and class balance degree, and the higher the value is, the more stable the clustering result is. The variation of different methods in multiple rounds of iterations is shown in Figure 6.

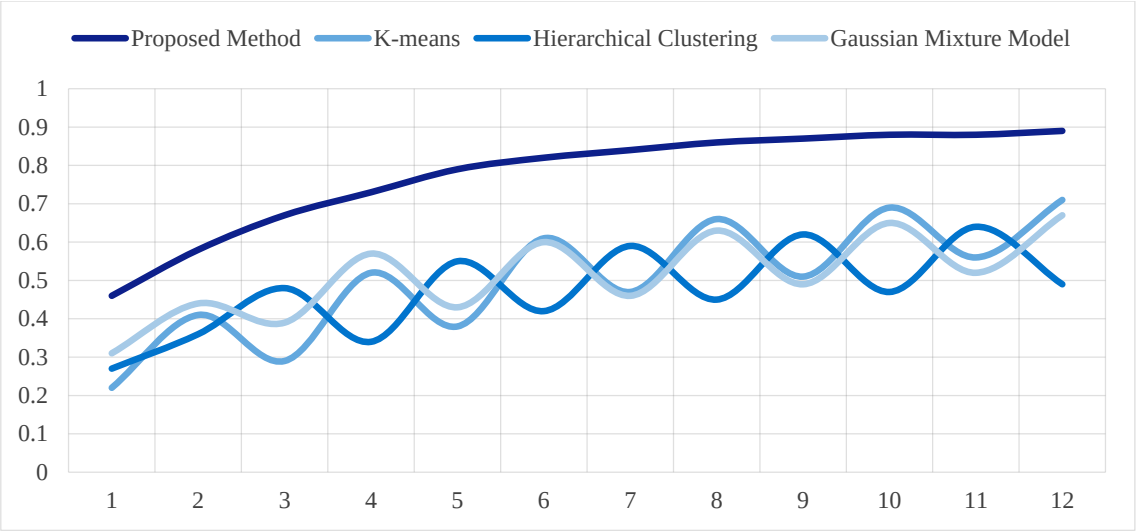

Figure 6: Structural evaluation value changes under multiple rounds of iterations.

It can be seen from Figure 6 that the fluctuation of the control method in the early iteration is large, especially in the batches with dense sample boundaries or similar post labels, the evaluation value is easy to fall down significantly. After the central sample screening and boundary sample correction of the proposed method were completed in the first four rounds, the structural evaluation value rose rapidly and remained in a high interval in the subsequent stages. Compared with the other three methods, although there are also periodic fluctuations in the curves of the proposed method, the overall trend is more stable and the amplitude contraction is more obvious in the later period. This shows that the adaptive local scale, boundary correction and group decision linkage mechanism can make the structure identification results tend to be stable faster, and also provide a more reliable clustering basis for the periodic update of the platform.

At the decision output level, we continue to count the changes in the accuracy of group decision making under multiple rounds of verification to investigate the stability of the model under continuous input updates. The accuracy variation of each method in multiple rounds of validation is shown in Figure 7.

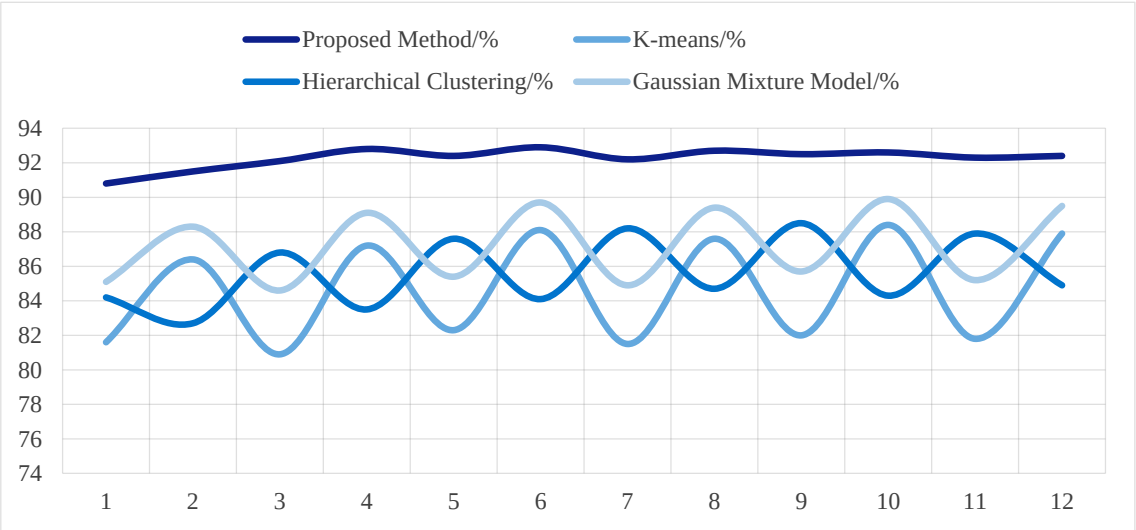

Figure 7: Group decision accuracy variation under multiple rounds of validation.

It can be seen from Figure 7 that the accuracy of the proposed method in each round always maintains a high level, with the lowest value being 90.8% and the highest value being 92.9%, and the overall fluctuation range is significantly smaller than that of the control method. K-means shows a large drop in multiple rounds, and although hierarchical clustering and Gaussian mixture models are more stable than K-means, they still have accuracy offset in batches with more boundary samples. The small fluctuation of the proposed method is directly related to the more accurate identification of central samples, the more adequate determination of boundary samples, and the joint modeling of post, region and ability information by comprehensive decision function. The results show that the model is not only effective in a single validation, but also able to adapt to the disturbance caused by the update of student behavior and the expansion of job data.

To further illustrate the contribution of each module to the clustering decision results, ablation experiments are conducted on the validation set. The complete model includes four parts: multi-source feature fusion, adaptive local scale density estimation, boundary sample correction and comprehensive decision mechanism. The ablation Settings were removed term by term, and the rest of the experimental conditions were kept constant to ensure the consistency of the comparison results. Recall is used to measure the detection ability of similar samples, Precision is used to reflect the accuracy of category assignment, F1 value is used to comprehensively describe the balance state of the two, and average response time is used to characterize the real-time feedback ability of the model in the platform environment. The ablation results are shown in Table 5.

Table 5: Comparison of ablation experiment results for each module.

Model Setting	Recall	Precision	F1-Score	Average Response Latency / s
Full Model	0.914	0.928	0.921	0.41
Without Multi-source Feature Fusion	0.883	0.897	0.890	0.37
Without Adaptive Local Scale	0.861	0.874	0.867	0.36
Without Boundary Sample Correction	0.892	0.904	0.898	0.39
Without Integrated Decision Mechanism	0.876	0.889	0.882	0.35

It can be seen from Table 5 that the full model achieves the optimal results in terms of Recall, Precision and F1 value. After removing multi-source feature fusion, the comprehensive differential expression of samples is weakened, and the three indicators are synchronously decreased. After removing the adaptive local scale, the F1 value decreases the most, which indicates that the dynamic adjustment of the local neighborhood range has a great impact on the stability of the clustering structure. After removing the boundary sample correction, the determination accuracy of the class transition region decreases. After removing the comprehensive decision-making mechanism, the discrimination ability of the clustering results in the position direction and regional output is weakened. Although the response delay of each ablation version is slightly lower, the performance degradation is more obvious, which indicates that the full model can obtain more stable clustering guidance results while maintaining acceptable computational overhead.

Combining Figures 4 to 7, it can be seen that the advantages of the proposed method are not only reflected in a single indicator, but in the whole link from feature construction, center identification to guided output. Multi-source feature fusion provides more stable input for clustering, density peak center identification enhances the clarity of category boundaries, and boundary sample correction and comprehensive decision-making mechanism further translate

this structural advantage into performance improvement in job guidance and regional recommendation.

5 Discussion

The advantages of the proposed model are mainly reflected in two aspects. Firstly, after unified coding of multi-source employment data, performance, practice, browsing, delivery and text semantics are put into the same feature space, and the clustering center is no longer affected by a single behavior frequency. Therefore, the contour coefficient is increased to 0.732, and the CH index is 418.6. On the other hand, the clustering results did not stop at the category division, but continued to enter the post matching, regional adaptation and confidence determination, so that the post guidance hit rate reached 89.7%, and the regional guidance consistency rate reached 91.3%. Existing density peak clustering studies mostly focus on distance rewriting, neighborhood construction and center screening, while job recommendation studies emphasize more on matching ranking. There is often a lack of direct connection between the two. In this paper, the clustering structure and the guidance output are connected into a calculation link, so that the boundary samples can also participate in the flexible decision, so that the clustering results are closer to the real employment choice of higher vocational college students. In multiple rounds of validation, the average accuracy of the model is 92.4%, and the fluctuation of the results is small, indicating that the model output has good stability. For the employment platform of colleges and universities, this modeling method not only supports the batch update of student samples, but also ADAPTS to the dynamic changes of the job database and continues to complete the calculation, so it has better adaptability in scenarios such as classification push, regional guidance and graduation destination tracking.

6 Conclusions

For the task of accurate analysis and guidance of employment intention of higher vocational college students, this paper constructs an analysis model based on density peak clustering, which connects multi-source employment data preprocessing, local density estimation, central sample identification, boundary sample correction and group decision output as a unified computing link. The experimental results show that the proposed method is based on a data set constructed by 4260 valid samples. On the validation set, the silhouette coefficient of the proposed method reaches 0.732, the CH index reaches 418.6, the accuracy of group decision-making reaches 92.4%, the hit rate of post guidance reaches 89.7%, the consistency rate of regional guidance reaches 91.3%, and the F1 value reaches 0.921. The average response time delay is controlled at 0.41 s, which indicates that the model can better identify the differences of students in position orientation, regional orientation and ability structure. The method in this paper still has some limitations. The sample sources are concentrated in the single-school data environment, and the professional structure and regional flow have obvious college characteristics. The effect of cross-college transfer still needs to be verified. Text features mainly rely on static semantic representation, and the description of short sentences, omitted expressions and phased intention fluctuations are not sufficient. At present, the guidance results are mainly based on position direction and region suggestions, and there is still room for expansion of the linkage modeling of enterprise position change, salary difference and long-term direction evolution. Subsequent research can be further carried out around the joint modeling of multi-school data, the introduction of temporal behavior characteristics, the update of dynamic job knowledge base and lightweight deployment, so as to enhance the generalization

ability, continuous computing ability and platform adaptation ability of the model, and provide more stable computing support for the tracking of graduation destination, classification push and fine employment service.

References

[1] Rasool Z, Zhou R, Chen L, et al. Index-based solutions for efficient density peak clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(5): 2212-2226. https://doi.org/10.1109/TKDE.2020.3004221

[2] Lu J, Zhao Y, Tan K L, et al. Distributed density peaks clustering revisited[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(8): 3714-3726. https://doi.org/10.1109/TKDE.2020.3034611

[3] Guo W, Wang W, Zhao S, et al. Density peak clustering with connectivity estimation[J]. Knowledge-Based Systems, 2022, 243: 108501. https://doi.org/10.1016/j.knosys.2022.108501

[4] Zhou Z, Si G, Sun H, et al. A robust clustering algorithm based on the identification of core points and KNN kernel density estimation[J]. Expert Systems with Applications, 2022, 195: 116573.https://doi.org/10.1016/j.eswa.2022.116573

[5] Zhu Y, Ting K M, Jin Y, et al. Hierarchical clustering that takes advantage of both density-peak and density-connectivity[J]. Information Systems, 2022, 103: 101871. https://doi.org/10.1016/j.is.2021.101871

[6] Ding S, Li C, Xu X, et al. A sampling-based density peaks clustering algorithm for large-scale data[J]. Pattern Recognition, 2023, 136: 109238. https://doi.org/10.1016/j.patcog.2022.109238

[7] Rasool Z, Aryal S, Bouadjenek M R, et al. Overcoming weaknesses of density peak clustering using a data-dependent similarity measure[J]. Pattern Recognition, 2023, 137: 109287.https://doi.org/10.1016/j.patcog.2022.109287

[8] Zhang Q, Dai Y, Wang G. Density peaks clustering based on balance density and connectivity[J]. Pattern Recognition, 2023, 134: 109052. https://doi.org/10.1016/j.patcog.2022.109052

[9] Zhao J, Wang G, Pan J S, et al. Density peaks clustering algorithm based on fuzzy and weighted shared neighbor for uneven density datasets[J]. Pattern Recognition, 2023, 139: 109406.https://doi.org/10.1016/j.patcog.2023.109406

[10] Xie J, Liu X, Wang M. SFKNN-DPC: Standard deviation weighted distance based density peak clustering algorithm[J]. Information Sciences, 2024, 653: 119788. https://doi.org/10.1016/j.ins.2023.119788

[11] Alsaif S A, Sassi Hidri M, Eleraky H A, et al. Learning-based matched representation system for job recommendation[J]. Computers, 2022, 11(11): 161. https://doi.org/10.3390/computers11110161

[12] Alsaif S A, Sassi Hidri M, Ferjani I, et al. NLP-based bi-directional recommendation system: Towards recommending jobs to job seekers and resumes to recruiters[J]. Big Data and Cognitive Computing, 2022, 6(4): 147.https://doi.org/10.3390/bdcc6040147

[13] González-Briones A, Chamoso P, Pavon J, et al. Job offers recommender system based on virtual organizations[J]. Expert Systems, 2024, 41(2): e13152. https://doi.org/10.1111/exsy.13152

[14] Mao Y, Cheng Y, Shi C. A job recommendation method based on attention layer scoring characteristics and tensor decomposition[J]. Applied Sciences, 2023, 13(16): 9464. https://doi.org/10.3390/app13169464

[15] Mao Y, Lin S, Cheng Y. A job recommendation model based on a two-layer attention mechanism[J]. Electronics, 2024, 13(3): 485. https://doi.org/10.3390/electronics13030485

[16] Al-Quhfa H, Mothana A, Aljbri A, et al. Enhancing talent recruitment in business intelligence systems: A comparative analysis of machine learning models[J]. Analytics, 2024, 3(3): 297-317.https://doi.org/10.3390/analytics3030017

[17] Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms[J]. Smart Learning Environments, 2022, 9(1): 11. https://doi.org/10.1186/s40561-022-00192-z

[18] Perez J C S, Manrique R F, Mariño O, et al. A course hybrid recommender system for limited information scenarios[J]. Journal of Educational Data Mining, 2022, 14(3): 162-188. https://doi.org/10.5281/zenodo.7304829

[19] Jena K K, Bhoi S K, Malik T K, et al. E-learning course recommender system using collaborative filtering models[J]. Electronics, 2022, 12(1): 157. https://doi.org/10.3390/electronics12010157

[20] Ahmad H K, Qi C, Wu Z, et al. ABiNE-CRS: course recommender system in online education using attributed bipartite network embedding[J]. Applied Intelligence, 2023, 53(4): 4665-4684.https://doi.org/10.1007/s10489-022-03758-z