Research on the Construction of Portrait of Higher Vocational College Students' Groups Based on Clustering Ensemble

Huihui Wang
Xuzhou Institute of Industrial Technology
wanghuihuiedu@163.com

Abstract. The rapid development of higher vocational education necessitates personalized teaching tailored to individual students. This paper presents an ensemble optimized clustering framework to classify higher vocational college students accurately. By integrating multiple clustering algorithms and employing optimization mechanisms, the framework enhances clustering quality. Experiments on real student data show that this method excels in clustering performance and adapts to various feature data types, yielding results aligned with actual situations. This research provides a data foundation for personalized education in higher vocational colleges, promoting educational equity and cultivating skilled talents.

Keywords: Higher Vocational Education; Clustering Ensemble; Clustering Optimization; Data Analysis.

1. Introduction

Higher vocational education in China has rapidly developed, crucial for cultivating skilled talents and supporting economic growth. However, the diverse academic foundations and personal backgrounds of students challenge the traditional "one-size-fits-all" model[1]. This study proposes an ensemble optimized clustering framework to accurately classify student groups, enabling tailored teaching programs. By leveraging data-driven analysis, this framework supports personalized education, promotes educational equity, and addresses resource shortages, thereby enhancing the overall effectiveness of higher vocational education.

2. Method for Constructing Portrait of Higher Vocational College Students' Groups Based on Clustering Ensemble

2.1 Data Collection and Preprocessing of Higher Vocational College Students

This study is based on data from 1000 students from a certain higher vocational college. Data were collected from multiple dimensions including personal basic information, family background, extracurricular activities both inside and outside the school, course grades, etc. Specifically, we collected data from various sources such as the school's academic management system, student personal files, etc., and integrated them to form the dataset as shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1 Example of Higher Vocational College Student Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ID</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Urban/Rural</td>
</tr>
<tr>
<td>Family Per Capita Income (10,000 RMB)</td>
</tr>
<tr>
<td>Math Score</td>
</tr>
<tr>
<td>Chinese Score</td>
</tr>
<tr>
<td>Participation in Club Activities</td>
</tr>
</tbody>
</table>

During data preprocessing, we handle missing values by filling them with the mean for numerical features and the mode for categorical features. Numerical features are standardized, and...
categorical features are One-Hot encoded[2]. The data is then split 80:20 into training and testing sets. Figure 1 shows the distribution patterns of numerical features in the training set.

![Figure 1 Distribution of Numerical Features in the Training Set](image)

2.2 Design of Clustering Ensemble Framework.

Our proposed clustering ensemble framework for higher vocational college student groups includes two modules: the base clustering component and the clustering ensemble component. The base clustering component utilizes algorithms like K-Means, hierarchical clustering, and DBSCAN, running independently on the training set[3]. The clustering ensemble component integrates these results using strategies such as voting or graph-based methods, followed by optimization techniques like simulated annealing. This approach combines multiple algorithms' strengths, allows modular design for easy extension, and uses optimization to improve clustering quality, yielding comprehensive and accurate results.

```python
# Base Clustering Component
base_clusterers = [KMeans(), AgglomerativeClustering(), DBSCAN()]
base_clustering_results = []
for base_clusterer in base_clusterers:
    base_clustering_results.append(base_clusterer.fit_predict(X_train))

# Clustering Ensemble Component
integrated_clusters = ensemble_clusterings(base_clustering_results,
                                          voting='evidence_accumulation')
optimized_clusters = optimize_clusters(integrated_clusters, X_train,
                                      optimizer='batch_search')
```

3. Experimental Study on Clustering Ensemble of Higher Vocational College Student Groups

3.1 Experimental Dataset.

In the data preprocessing stage, we first utilized Word2Vec to vectorize students' self-description texts, transforming textual features into 100-dimensional vectors. Next, numerical features such as age and family income were standardized, while categorical features such as gender and urban/rural status were encoded using One-Hot encoding. The final dataset inputted into the model consists of 234 features, including a 100-dimensional text vector feature[4].

3.2 Baseline Models

To evaluate the proposed clustering ensemble framework, three baseline models are used for comparison:1)K-Means Clustering: A classic partitioning algorithm, optimized for the Silhouette Coefficient, with K ranging from 3 to 10.2)Hierarchical Clustering: A bottom-up approach using single-linkage, complete-linkage, and average-linkage, selecting the best result based on the Clustering Quality Index (CQI).3)DBSCAN: A density-based algorithm, iterating over eps and min_samples parameters to achieve the highest Adjusted Rand Index (ARI)[5]. Table 2 shows the performance of these models, illustrating the need for the clustering ensemble. The next section compares these baselines with the proposed framework.
### 3.3 Baseline Models

#### 3.3.1 Analysis of Model Parameter Effects

In this experiment, the basic clustering components employed three algorithms: K-Means++, Weighted Average Clustering, and HDBSCAN. These algorithms were independently run on preprocessed training data to generate initial clustering results. Firstly, we analyzed the impact of the number of basic clustering classifiers on the final ensemble performance[6]. The overall clustering quality metrics reached optimal levels when 7 basic classifiers were used, as shown in Figure 2. Using too many or too few basic classifiers resulted in decreased performance; excessive classifiers could introduce redundant information, while too few could fail to provide sufficient diverse clustering perspectives.

![Figure 2 Impact of the Number of Base Clusterers](image)

Next, we investigated the impact of different ensemble strategies on model performance. In this experiment, we tried three strategies: simple voting, graph-based integration, and evidence accumulation-based integration[7]. Table 3 shows the evaluation results on the test set, indicating that the evidence accumulation-based integration strategy performs the best.

<table>
<thead>
<tr>
<th>Ensemble Strategy</th>
<th>ARI</th>
<th>Silhouette</th>
<th>Kendall's τ</th>
<th>CQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Voting</td>
<td>0.725</td>
<td>0.691</td>
<td>0.782</td>
<td>39.21</td>
</tr>
<tr>
<td>Graph-Based Integration</td>
<td>0.743</td>
<td>0.718</td>
<td>0.805</td>
<td>41.35</td>
</tr>
<tr>
<td>Evidence Accumulation-Based</td>
<td>0.768</td>
<td>0.739</td>
<td>0.827</td>
<td>43.17</td>
</tr>
</tbody>
</table>

Finally, we compared optimization algorithms, including batch search, simulated annealing, and particle swarm optimization. The results showed that the batch search algorithm balances clustering quality and computational efficiency well, as shown in Table 4.

<table>
<thead>
<tr>
<th>Optimization Algorithm</th>
<th>ARI</th>
<th>Time Spent (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch Search</td>
<td>0.768</td>
<td>235</td>
</tr>
<tr>
<td>Simulated Annealing</td>
<td>0.761</td>
<td>478</td>
</tr>
<tr>
<td>Particle Swarm Optimization</td>
<td>0.755</td>
<td>619</td>
</tr>
</tbody>
</table>

Based on the above analysis, we have determined the optimal parameter configuration for the clustering ensemble framework: the number of base clusterers is 7, the ensemble strategy is evidence accumulation-based integration, and the optimization algorithm is batch search. In the subsequent sections, we will further evaluate the model performance under this parameter configuration[8].

#### 3.3.2 Analysis of Model Parameter Effects

After determining the optimal model parameters, we conducted a comprehensive evaluation of the clustering ensemble model's performance on the test set. As shown in Table 5, the model
advances in education, humanities and social science research

ISSN: 2790-167X

Volume-11-(2024)

achieved the best performance on all four evaluation metrics, significantly outperforming the baseline models introduced in Section 4.3.

Table 5 Performance Comparison between Clustering Ensemble Model and Baseline Models

<table>
<thead>
<tr>
<th>Model</th>
<th>ARI</th>
<th>Silhouette</th>
<th>Kendall's $\tau$</th>
<th>CQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>0.637</td>
<td>0.695</td>
<td>0.756</td>
<td>26.83</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>0.571</td>
<td>0.619</td>
<td>0.697</td>
<td>22.75</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>0.682</td>
<td>0.559</td>
<td>0.638</td>
<td>30.12</td>
</tr>
<tr>
<td>Clustering Ensemble</td>
<td>0.768</td>
<td>0.739</td>
<td>0.827</td>
<td>43.17</td>
</tr>
</tbody>
</table>

The clustering ensemble model achieves an ARI of 0.768, indicating high consistency with true labels. Its Silhouette value of 0.739 reflects strong intra-cluster cohesion and inter-cluster separation[9]. With a Kendall's $\tau$ of 0.827 and a CQI of 43.17, the model's effectiveness is validated. Feature distribution analysis, shown in Figure 3, reveals that one student group has high mathematics scores, moderate family income, and low extracurricular activity participation. This information aids teachers in understanding group characteristics and tailoring educational strategies accordingly.

3.3.3 Visualization Analysis

We performed a visualization analysis of the clustering ensemble model's output using the t-SNE technique to reduce high-dimensional data to two dimensions. Figure 4 shows distinct separation boundaries between clusters and high concentration within clusters, indicating good clustering quality. Figure 5 labels specific sample points from Figure 4, highlighting primary feature values. This visualization reveals distribution differences between groups and provides insights into subtle group features, offering crucial support for informed educational decisions[10].

3.4 Comparison Analysis with Baseline Models

To evaluate the advantages of the proposed clustering ensemble framework, we compared it with three classical baseline models (K-Means, hierarchical clustering, and DBSCAN) on the same test
set. Table 8 in Section 4.4.2 shows that the clustering ensemble model outperforms the baselines on ARI, Silhouette, Kendall's τ, and CQI metrics. Figure 6, a radar chart, visually highlights the ensemble model's superior performance across all metrics, confirming the effectiveness of integrating multiple base clusterers and optimization strategies.

![Figure 6 Performance Comparison of Different Models](image)

4. Conclusion

This paper presents an optimized clustering framework for classifying student groups in vocational education. By leveraging multiple clustering algorithms and optimization strategies, the framework enhances clustering quality. Experiments on real vocational student data show excellent performance across various metrics and adaptability to different feature types. Visual analysis confirms the model's superior performance. This research supports data-driven personalized education in vocational colleges, promoting educational equity and talent development.
References


