Lab Report-CW3

kaijie.Yang, 2256526 TAs name: Yiqiang Cai, Xin Gao, Changwei Li

May 22, 2024

1 Introduction

In today's information age, the amount of data is increasing exponentially, and how to extract useful information from massive data has become an important challenge. As an unsupervised learning method, cluster analysis is widely used in data mining and machine learning. By dividing data into different groups (clusters), clustering models can reveal the internal structure in the data and help us understand the laws and relationships behind the data. Through cluster analysis, we can find common characteristics and nuances between students' grades and majors.

In this experiment, three different clustering methods are compared with original classification labels by adjusting their parameters and clustering criteria. Therefore, it is found whether it is a suitable clustering model and clustering method. In the process of testing, cross-validation and cluster center visualization are used to show the model training results in detail and clearly, so as to obtain the specific relationship between student achievement and major.

2 Methods

2.1 Gaussian Mixture Model

Gaussian mixture model (GMM) is a widely used clustering algorithm in statistics and data mining. GMM assumes that all data points are generated by clusters with Gaussian distributions. In data preprocessing, StandardScaler model is used, so that the mean value of the data is 0 and the variance is 1. In terms of feature selection, the two features of Index and Gender are removed, so that the content contained in the data is more suitable for the differentiation of students' majors. For the following two models, the same processing method is also carried out, and if there is a difference, it will be specified.

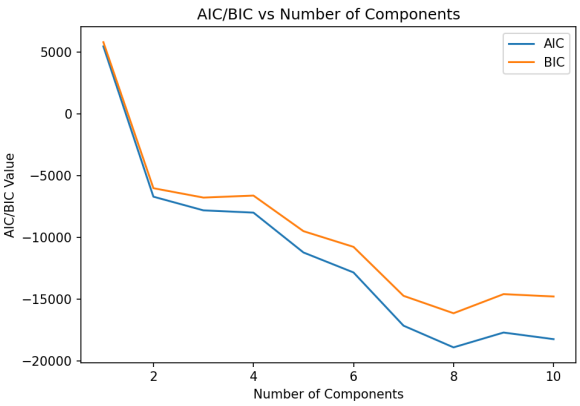

Figure 1. AIC and BIC value curves of GMM model

In the process of applying gmm model, the AIC and BIC values of the model are calculated for the number of components in the model. Find out how many classes the model thinks the data set should be divided into, as shown in **(figure 1)**.

It can be seen from the value curves of AIC and BIC that the model clustering effect is the best when the number of components is 8. In the scatter plot presented next, the clustering effect of the number of components 4 and 8 will be compared.

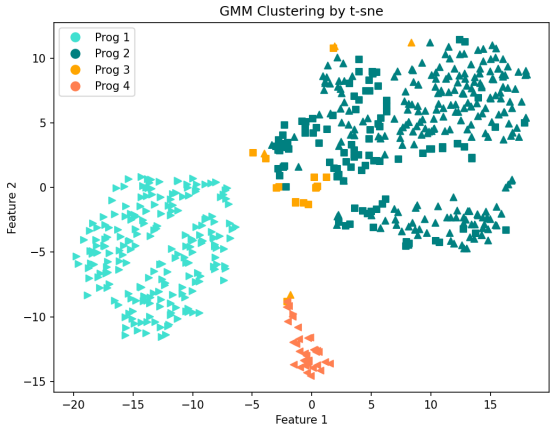

Figure 2. Visualization of gmm clustering model for t-sne scatter plots

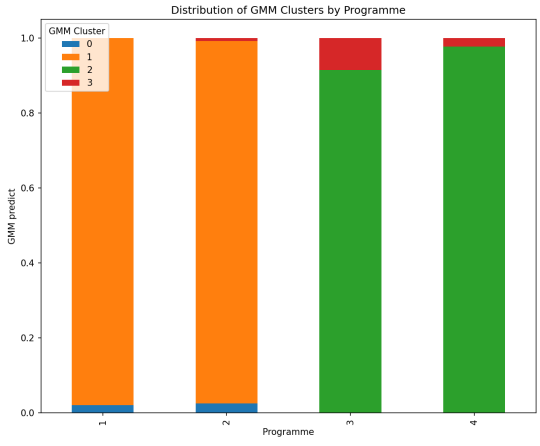

Figure 3. The cross-tabulation of the gmm visualization model clearly shows the comparison of the predicted data with the original data

For the demonstration of the clustering performance of gmm models, the dimensionality reduction methods of pca and t-sne are considered. After observation, t-sne has the best dimensionality reduction performance, and the data point distribution is clearly

visible. At the same time, for the judgment of classification effect, we use the representation of colors and graphs, different colors represent the original student professional data, and different graphs represent the distribution of different clustering results **(figure 2)**. It is obvious that programme1 and programme2 students are successfully separated by the clustering model. However, data points of different shapes in other color blocks occupy a large proportion, which indicates that the clustering effect of other programmes is not good.

At the same time, the cross graph of gmm models shows the proportion of model clustering results in real labels for the original programme, as shown in **figure 3**. In general, the gmm clustering effect is slightly effective, clustering out the first two labels, but the model inferred that this data set can be divided into 8 clusters, which is a certain gap for the real actual situation.

2.2 K-means

K-means is a commonly used clustering algorithm that aims to divide data points into K clusters so that each data point belongs to the center point of the nearest cluster. The algorithm is based on iterative optimization, which realizes the clustering process by alternately updating cluster centers and redistributing data points. In the experiment, the elbow rule and contour coefficients were calculated to determine the k value accepted by the model, as shown in **(figure 4)**.

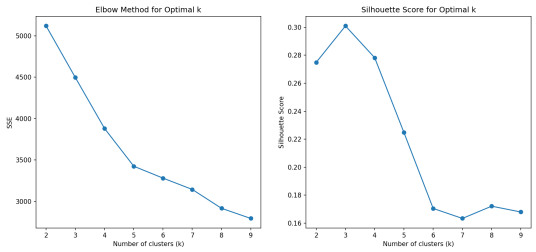

Figure 4. Elbow rule and contour coefficients with corresponding k values

The K-means model is different from the gmm model in the process of training, which specifies the number of cluster components to show the difference between the clusterer and the real label judgment. The former pays more attention to the clustering ideas of the model itself, whose k value is determined by the specific data set, and the clustering effect is automatically determined by the model. In this way, we can clearly see the cluster type between the data and find out the relationship between the feature and the cluster center. The model automatically selects the k value as 8. Then 8 will be used as the number of clustering centers of k-means to train the model, and different colors will be used to determine the clustering results of k-means in the visual scatter plot with t-sne reduction **(Figure 5)**.

At the same time, the process of cross-validation was added to the drawing of the crosstab, and the data clustering was completed automatically by the model, which made the model classification and clustering results more convincing **(Figure 6)**. It can be seen from the crosstab that only programme3 in the clustering result matches the actual label. Moreover, it can be found in the data points of the scatter plot that k-means does mark the category of each student by Euclidean distance, and there is a clear clustering result.

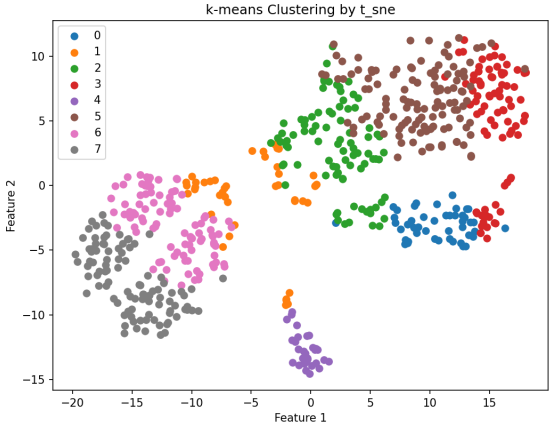

Figure 5. k-means visual cluster map, different colors represent different model clustering results

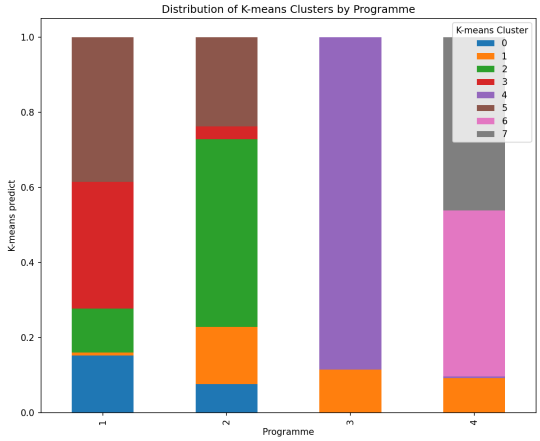

Figure 6. What labels do k-means crosstabs that represent clustering results actually belong to

2.3 Hierarchical Clustering

Compared with other clustering methods, hierarchical clustering can better reflect the similarity relationship between data, and is more suitable for interpreting clustering results and processing small sample data. At the same time, compared with K-Means and other methods that need to specify the number of clusters in advance, hierarchical clustering does not need to specify the number of clusters in advance. It can automatically build the hierarchy of clusters until a single cluster containing all the data points is formed. It can clearly show the correlation between the data. For this data set, the hierarchical clustering result tree is

shown in **Figure 7**.

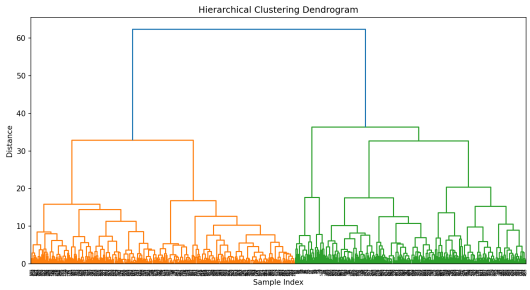

Figure 7. Tree diagram of hierarchical clustering

The tree diagram shows the cluster structure of data points at different levels, starting at the bottom, where each data point is initially treated as a separate cluster and then gradually merged into larger clusters until all data points are merged into a single cluster. It shows the similar structure between each data point, finds out the similarity between points, judges their similarity step by step, and gathers into the same cluster. Finally, it is manifested in two kinds of clustering.

The next diagram we present is a scatter plot with cluster center 4 manually controlled and a crosstab to see how well the hierarchical clustering fits the data label given the component **(Figure 8)(Figure 9)**.

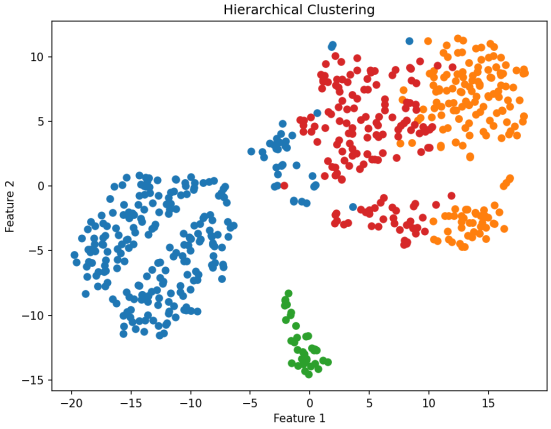

Figure 8. Clustering results of hierarchical classification

3 Conclusion

In general, the clustering function for this dataset is not as good as ideal. But we can change the way we think about making the clustering tag accurate to the actual tag. It is equivalent to making the model cluster the class centers that meet the specified label features, which needs to add the role of supervised learning in the training process of the clustering model, and improve its accuracy rate as the evaluation standard. In the process, the

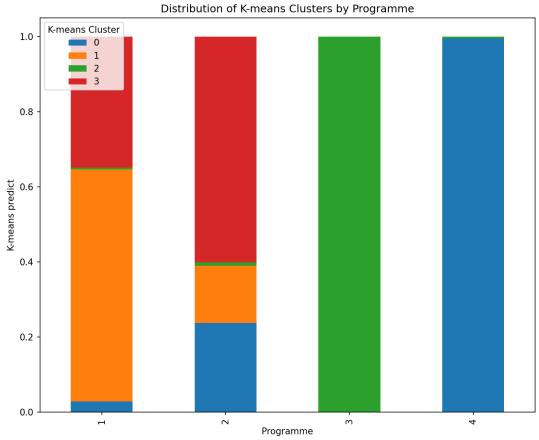

Figure 9. A crosstab of hierarchical classification

network cross-validation is carried out to obtain the best model clustering parameters, and then the specific parameters are used to operate the original data to get clear and clear clustering results.

The whole experiment tends to get a comparison between the cluster and the original label, which helps a lot in the way mentioned above. The specific clustering effect depends on the parameter adjustment result of the model running. If the above method is not adopted, it is difficult for the cluster label to correspond to the real label.