Lab Report of Coursework3

Data Cluster Methods

Hongkun Jiang, 2253854
TA : Yiqiang Cai

***Abstract*—For the purpose of extracting the feature of the data and evaluate how much information it contains with program, the experiment applies cluster and compare the result. Without the program information answer given, unsupervised learning methods mean to group the data based on distance between points.**

***Index Terms*—unsupervised learning, Gaussian Mixture Model, K-means, hierarchical clustering, etc.**

I. Introduction

AS the lab required to observe the data and visually distinguish the distribution of programs, as well as furthermore generating classifiers to gain accuracy results, here comes the new evaluation of group to form clusters based on the internal distance and estimate the relationship between clusters and programs.

The following unsupervised learning algorithms will be applied. Firstly, use GMM to reflect the distribution of raw data, trying to fit the same clusters as the actual program label. For the next step, observe and alter the n-components of k-means and estimate which is the best amount of clusters. Eventually, by comparing the different varieties of hierarchical cluster conclude the probable classification structure of data points.

II. Methodology

In the following experiments, the dataset will be spilt into a training set and a test set, which represent 75% and 25% respectively. The training part is used for clustering. In this stage, coefficient such as Silhouette Scores is referenced to select a better hyper-parameter for better fitting clusters. Nevertheless, it is normally hard to detect whether a better cluster model has to do with the real distribution of program. The only aim is to attempt to improve the cluster itself based on the number of program which is 4. Connecting the clusters with specific program types, we will conclude accuracy and may analyse to what extend the cluster is relevant to program.

A. Gaussian Mixture Model

By the random forest classifier we have the importance ranking of features of which implies more information about program, which is also similar to the result of information gain. Therefore, index and gender are excluded. Firstly, a visual comparison is plotted to determine if the default cluster amount 4 matches the actual case.

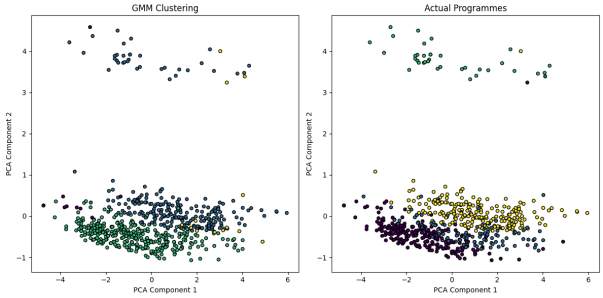

Fig. 1. GMM 4n visual

As shown in the comparison figure, three types of programs clearly match the actual case while one particular program shows no tendency. Now it is a try to calculate the result.

```
(0.748,
 array([[52,  0,  0, 10],
        [17,  0,  0, 12],
        [ 0,  0,  6,  0],
        [ 0,  0,  0, 58]]))
```

Fig. 2. GMM 4n result

Such accuracy is unexpectedly high with no program 2 is classified as predicted. The experiment above, as clustering, performs even better than classification. This is probably because the feature ”program” it self is added in, which carries plenty of information about itself. To determine how much can the other features reflect, the program is dropped.

```
(0.555,
 array([[45,  0,  0, 17],
        [ 9,  0,  0, 20],
        [ 6,  0,  0,  0],
        [17,  0,  0, 41]]))
```

Fig. 3. GMM result without program

The accuracy remains the level of 0.555, which seems to be much higher than the nature guess of 0.25. According to the confusion matrix, program 2 and 3 are totally wrong and divided to program 1 and 4. Therefore, one suppose is that the data are splitted into only two valid groups due to distinct gap of some programs and result in a high overall accuracy. Concerning this reason, we may try to optimize the number of clusters.

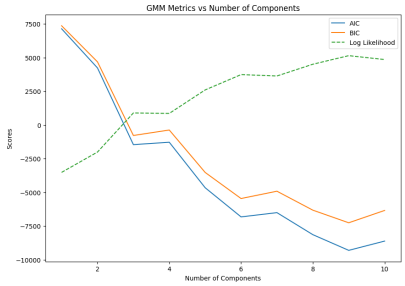

Fig. 4. GMM Metric

The standard above in the graph assess the tightness and performance of cluster itself, which do not essentially connected to program information but might cluster better. To select a point with low AIC, BIC and high likelihood, 6 and 9 is outstanding. However, 9 clusters is much bigger than 4 programs and may cause overfitting, 6 is attempted.

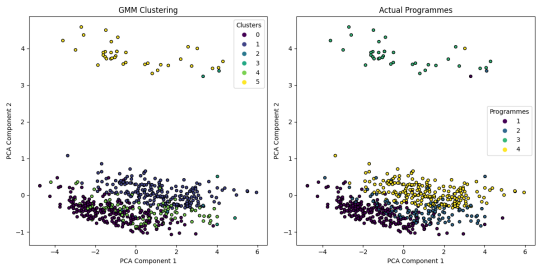

Fig. 5. GMM 6n visual

```
(0.5870967741935483,
 array([[48, 10,  0,  4],
        [ 0, 20,  0,  9],
        [ 0,  0,  6,  0],
        [ 0, 41,  0, 17]]))
```

Fig. 6. GMM 6N result

In this experiment, the program feature is taken into consideration. According to the PCA graph, to begin with, one of the cluster has include a bunch of points of program 2 and others clusters still state a tendency to be relevant to one specific programs. With lower accuracy of 0.587, the clustering method performs unsatisfactory in contrast with cluster amount of 4. However, the confusion matrix corresponding to 6 clusters claims that the program type 2 are successfully identified, though partly. As for the other programs, the majority of them can still be correctly classified.

As a short summary of GMM, the test set prove that the clustering of train set can somewhat reflect the information of program feature while being not so effective as classification. Similarly, it do not perform well when it comes to program 2.

B. K-means Clustering

Having explored the possible amount of information implied by the cluster via GMM, an issue that program 2 is difficult to divided from program 1 and 4 (especially 4 actually) is identified from clustering and most cases of classification To give a further explain of this from the perspective of clustering, here applies the k-means algorithm. The K-means method can be regarded as a specific case of GMM. It focus on the character of the clusters.

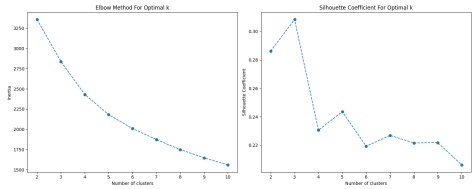

Fig. 7. kmeans standard

There are elbow plot and line charts of silhouette coefficient of k-means clustering. The first figure shows a sharp decrease of inertia when the number reaches 4 or 5. In contrast, silhouette coefficient reaches the peak while the number equals to 3 and dramatically drop when becomes four.

Theoretically, these two parameter are supposed to corresponding to each other in one or several clusters, while here having huge disagreement choosing 3 and 4. Visual graphs are plotted for clarification.

Fig. 8. kmeans 3 cluster

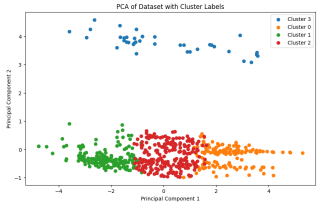

Fig. 9. kmeans 4 cluster

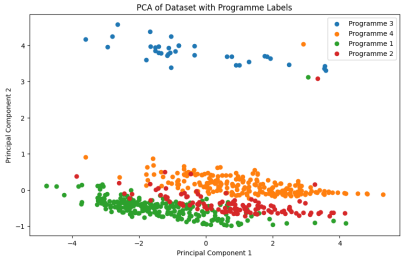

Fig. 10. kmeans actual program

Three clusters essentially group by three types of programsthough exist deviation. However, as the components grow to 4it present differently as the actual program distribution. Consequently, implementing 3 clusters seems to be more reasonable.

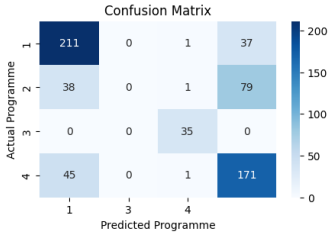

Fig. 11. kmeans confusion matrix

As an explanation of the divergencevalue in elbow plot only focus on the internal distances of each cluster. Meanwhile, the silhouette coefficient takes into account both the tightness and the separation of the clusterswhich may mean that while four clusters may be a possible outcome, three clusters are the easiest to distinguish at the macro level. The confusion matrix of 3 cluster also demonstrate the same result that program 2 is not validly identified.

C. hierarchical clustering

In the section above, the experiment do not achieve to separate the program 2 from program 1 and 4. Hierarchical clustering may act as an effective algorithm since it normally clustered based on similarity from the bottom. The number of cluster is set as fixed number 4.

Having attempted various sorts of linkage of hierarchical clustering, the ward seems to be the best. Outcome will be displayed as follow.

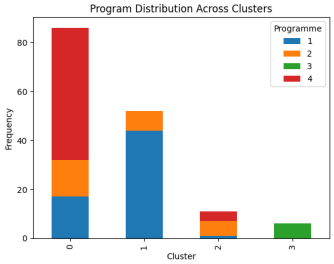

Fig. 12. program distribution histogram

Confusion Matrix:

44	1	0	17
8	6	0	15
0	0	6	0
0	4	0	54

Fig. 13. HC confusion matrix

The histogram is a preliminary observation of program distribution across the clusters. The image exhibits that within each cluster there is a program that occupy the prior proportion. This may display an effective output when calculating. The confusion matrix which is concluded from the test set show favorable effects and produce positive outcomes.

Nonetheless, the precise accuracy of program 2 remains the level of natural guess. The rationale behind this is, firstly, its feature is highly overlap with its adjacent program. Secondly, the number of students are less and the cluster is easily to be affected by impurity.

III. Discussion and Conclusion

Under the condition that index and gender are excluded and the rest feature are adopted, the performance of clustering indicates that distribution of data confirm the program information and similar to the classification outcomes. In other words, the clusters can connect with certain programs.

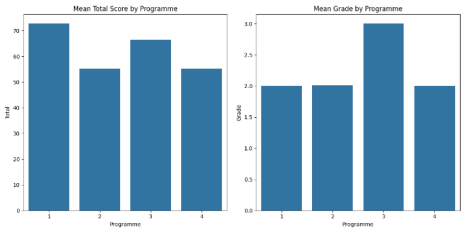

Fig. 14. mean score

In general, the program 3 is always clearly divided. This is firstly discovered in CW1 when adding the feature ”Grade”. According to personal analysis, this is because the Grade of program3 students are all 3 and other students are grade 2. So it is possible to 100% tell program 3. Furthermore, program 1 scores much more than program 2 and 4 in total, while 2 and 4 are approximately the same and thus hard to classify. The findings via clustering match it because in a vertical dimension which the mapping of grade program 3 are faraway from others and in horizontal axis which might be mapping of a type of score program 1 tends to be different from program 2 and 4 while these two are mixed.