Lab Report For INT104 Coursework 3

Xu Chen
2257453
TA: Changwei Li

***Abstract*—Unsupervised learning is an important part of traditional machine learning. This report is based on a data set of INT104 student final exam results from the previous year, and compares Gaussian Mixture Model , k means , hierarchical clustering for data classification. In addition, we map these models to two dimensions and use tsne dimensionality reduction technology to visualize the hyperplane of the model, so as to better understand how to select models and extract features under classification problems, and explain the causes of phenomena based on theory and experiment.**

I. Introduction

In this report, our goal is to use traditional unsupervised learning to complete four classification tasks based on index, gender, grade, total, MCQ, Q1, Q2, Q3, Q4 and Q5, with a sample size of 619, to distinguish which students belong to the programme .

In this article, all methods below will perform data normalization operations in advance

$$x' = \frac{x - \text{mean}}{\text{std}} \tag{1}$$

#	Feature	Average Mutual Information
1	Grade	0.198751
2	Total	0.187396
3	MCQ	0.104358
5	Q2	0.087815
7	Q4	0.086718
8	Q5	0.058307
6	Q3	0.048980
4	Q1	0.040192
0	Gender	0.032749

TABLE I
The Importance of Feature Based on Average Mutual Information

Table 1 shows me using the average of 10 groups of random numbers to calculate the mutual information of each feature and item. The higher the ranking, the higher the importance. This is an important basis for how I will extract different features to calculate the accuracy. I will directly select the top Four as initial features, and then incremented one by one.

Figure 1 mainly explains why we later use tsne dimensionality reduction technology as observation instead of other dimensionality reduction techniques. It can be seen from PCA that the data points are very close together, which is not conducive to hyperplane drawing.(Related parameters in tsnecomponents=2, random state=60, perplexity=100, learning rate=50, iteration =1000, early exaggeration =70)

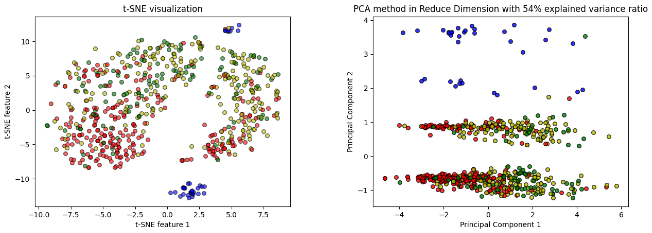

Fig. 1. Comparison of PCA and TSNE in two dimensions

Not only that, since we classify the samples into four categories but we don’t know which category corresponds to the corresponding programme, here we use the Hungarian algorithm for matching to ensure that the most ture labels are predicted.

It is worth noting that I have drawn the model boundary in subsequent figures, but this boundary is fitted with the help of the knn algorithm. The reason is that tsne is a nonlinear dimensionality reduction technology and cannot achieve the reverse transformation from point to line. In addition, in the dimensionality reduction graph, I keep the data on the unextracted features for dimensionality reduction, but the model extracts the features and then maps them to the graph to prevent the different dimensions of the training from interfering with the prediction. This is because unsupervised learning has a huge impact on the dimensions and data. Extremely sensitive to sparsity. In the following dimensionality reduction diagram, the scatter points red, green, blue, and yellow correspond to 1, 2, 3, and 4 respectively. The shadow of the decision boundary does not correspond to the scatter points on the diagram.

II. Gaussian Mixture Model

TABLE II
Classification of Gaussian Mixture Model under different features

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.533	0.526
Grade, Total, MCQ, Q2 ,Q4	0.249	0.554
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.430	0.199
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.307	0.537
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.333	0.505
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.290	0.150

Table 2 shows that when Grade, Total, mcq, and q2 are features, the GMM model has the highest accuracy. We extract these as features and draw the confusion matrix of all the data to obtain Figure 2. Figure 3 shows the difference between prediction and actual.

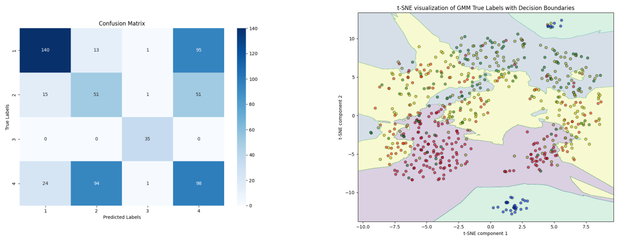

Fig. 2. Actual performance of Gaussian Mixture Model

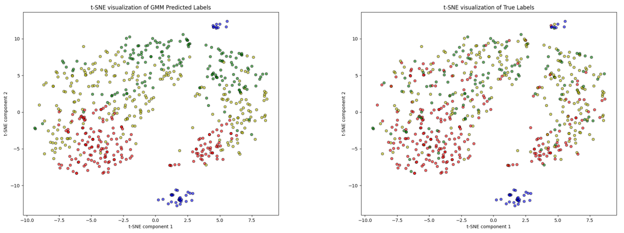

Fig. 3. Comparison between true label and predicted label in GMM

III. K-Means

TABLE III
Classification of K-Means under different features

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.519	0.505
Grade, Total, MCQ, Q2 ,Q4	0.529	0.516
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.494	0.500
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.478	0.526
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.478	0.538
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.461	0.413

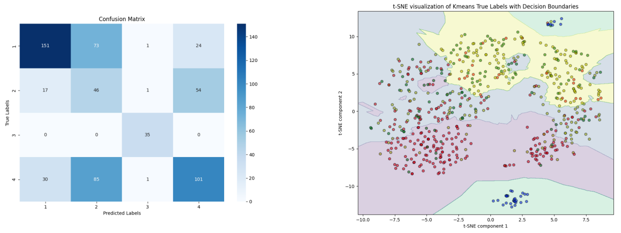

Fig. 4. Actual performance of K-Means

Table 3 shows that when Grade, Total, mcq, q2, and q4 are features, the k-means model has the highest accuracy. We extract these as features and draw the confusion matrix of all the data to obtain Figure 4. Comparing Figure 4 and Figure 2, we can see that they all learned Category 1 very well. The corresponding red part in the decision boundary diagram.

IV. Hierarchical Clustering

Table 4 shows that when Grade, Total, mcq, q2, q4, and q5 are features, the Hierarchical Clustering model has the highest accuracy. We extract these as features to draw the confusion matrix of all data to obtain Figure 4. We can see from the comparison of the color depth on the diagonal in Figure 2, Figure 4, Figure 6 that this model and test are the best. The reason may be that tsne itself likes to put similar position distances close together to cater to this algorithm.

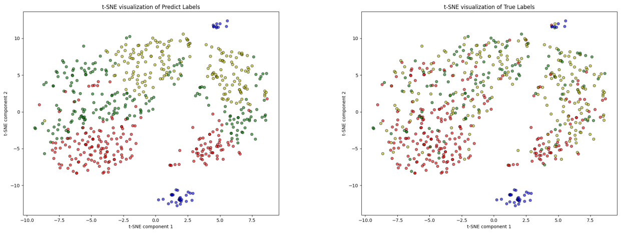

Fig. 5. Comparison between true label and predicted label in K-Means

TABLE IV
Classification of Hierarchical Clustering under different features

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.536	0.500
Grade, Total, MCQ, Q2 ,Q4	0.522	0.559
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.558	0.564
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.505	0.553
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.503	0.569
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.503	0.559

V. Ensemble Learning

Table 5 compares the accuracy of the best models. Among them, the ensemble uses hard voting and is composed of the above three and the features extracted by each model are different. From the table, Hierarchical Clustering performs best, but I think this is a problem of data volume. , from the integrated decision boundary, I think his generalization performance is the best.

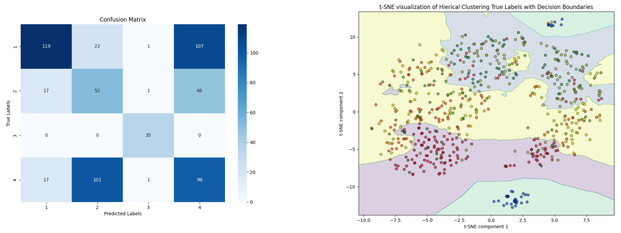

Fig. 6. Actual performance of Hierarchical Clustering

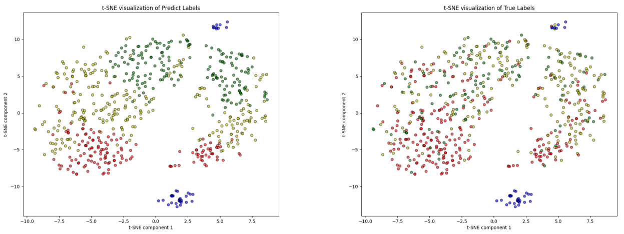

Fig. 7. Comparison between true label and predicted label in Hierarchical Clustering

TABLE V
CLASSIFICATION OF CURRENT STATE OF ART MODELS UNDER DIFFERENT FEATURES

Model with specific feature	Cross-validation accuracy	Test accuracy
GMM (Grade, Total, MCQ, Q2)	0.533	0.526
K-Means (Grade, Total, MCQ, Q2 ,Q4)	0.529	0.516
Hierarchical Clustering (Grade, Total, MCQ, Q2 ,Q4 ,Q5)	0.558	0.564
Ensemble	0.520	0.510

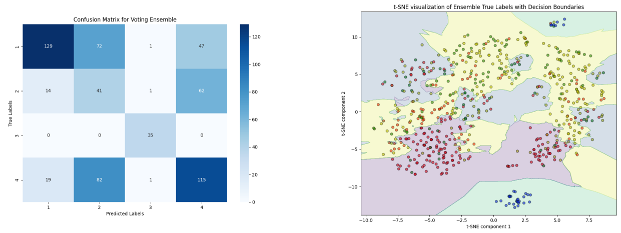

Fig. 8. Actual performance of Ensemble Learning

VI. CONCLUSION AND DISCUSSION

In this report , we compares Gaussian Mixture Model , k means , hierarchical clustering for data classification. In addition, we map these models to two dimensions and use tsne dimensionality reduction technology to visualize the hyperplane of the model, so as to better understand how to select models and extract features under classification problems, and explain the causes of phenomena based on theory and experiment.

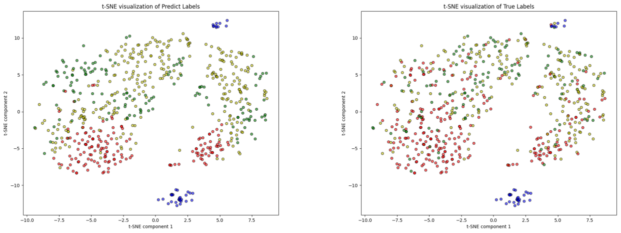

Fig. 9. Comparison between true label and predicted label in Ensemble Learning