Lab Report of Coursework1

PCA application for classifying data with different features

***Abstract*—This is a lab report of INT104 Artificial Intelligence coursework 1. The core task of this coursework is to classify the data with one specific feature difference by Principle Component Analysis (PCA). By observing the raw data and processing the data, suitable data will be selected to perform PCA The final result is a scatter plot with the best classification student can achieve by applying PCA.**

***Index Terms*—data process, PCA, box plot, INT104, etc.**

I. Introduction

PRoviding a file that consist of the final marks and other basic information of students in a module, who are from different grades and programs, the coursework requires to ignore the column program in the file and use the remaining data to distinguish the students in different programs. Data observation and PCA scatter plot are demanded in this task. Firstly, students are supposed to observe the feature of the raw data in a box plot. Afterwards, the data is scaled and the distribution of in new box plot is shown and appropriate features are chosen. After applying PCA to these columns, a scatter plot with default principal components (PCs) is stated to show the classification. Attempts of alternative PCs to draw scatter plots will be based on the box plot of the PCs. Meanwhile, other combinations of columns will be tried. Eventually, the combination that fits the classification of program most will be adapted.

II. Methodology

To begin with, the file is read by python. The coursework data has 11 columns in total, including 618 students. Columns of basic information are index, gender, program and grade. The rest of columns are the final mark components including total marks, MCQ and scores from Q1 to Q5. The box plot of raw data is shown as follow.

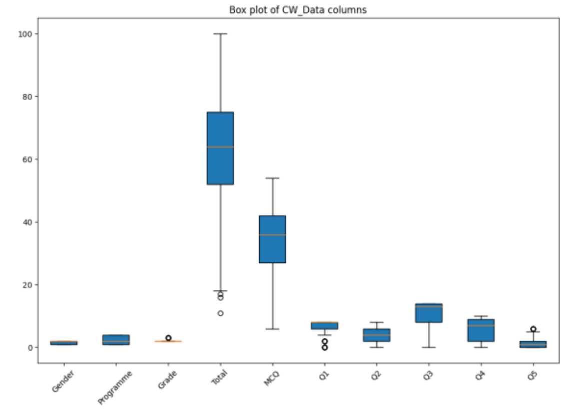

Fig. 1. the box plot of raw data without index

As is shown in figure 1, the box plot of raw data shows an enormous gap between the range of different columns. This character is decided by the real data, which do not mean some of the features are not suitable for data analyse. As an explanation, for example, the range of gender and grade are binary and program varies from 1 to 4 However, gender do not make sense to the grade, but grade do. Consequently, the box plot of scaled data is drawn.

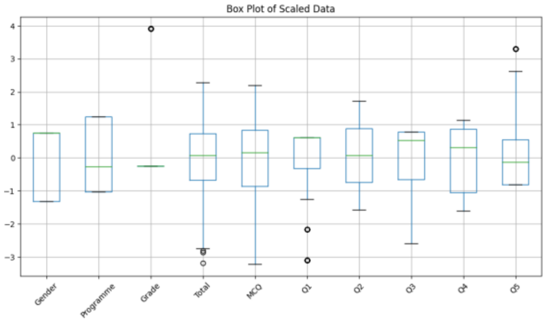

Fig. 2. the box plot of scaled data without index

The scaled data has used the method of standardization. It minimizes the natural range difference of the columns. Accordingly, the value is put in a similar range. The gender is binary and thus the value is extreme and gender itself is considered meaningless about scores. Program must be drop since it is question itself. In addition, the grade seems to be less important. Therefore, the rest stuffs are chosen to apply PCA.

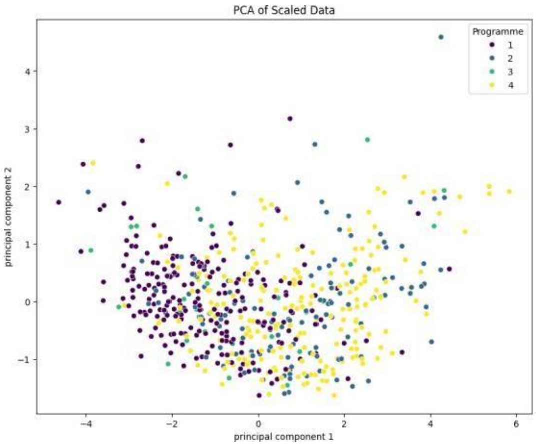

Fig. 3. scatter plot 1.0

Applying PCA actually means find two new coordinate axes by rotation and shift the default x and y axes. It is usually calculated according to the value of variance. The scatter plot, in fact, is the mapped raw data in the new coordinate system. The PC1 and PC2 are selected acquiescently due to the greatest variance they have.

In the graph above, different program shows different colors to observe. However, this combination with the two default principal components does not seem to be ideal to classify. PCA is used to classify the mapped data scatter and do not have to consider variance so other PCs are considered.

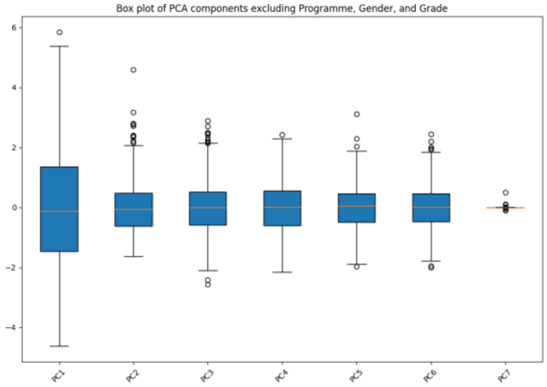

Fig. 4. box plot of PCA 1

The scatter plot of PC 1 and PC 2 has a poor tendency to classify all the 4 programs. Only 1 and 4 can be distinguished. In a box plot, quartile range, outliers and median among the rectangle are main factors to assess the reliability of a element. According to the box plot, PC 4 and PC 1 are selected due to fewer outliers and more concentrated data.

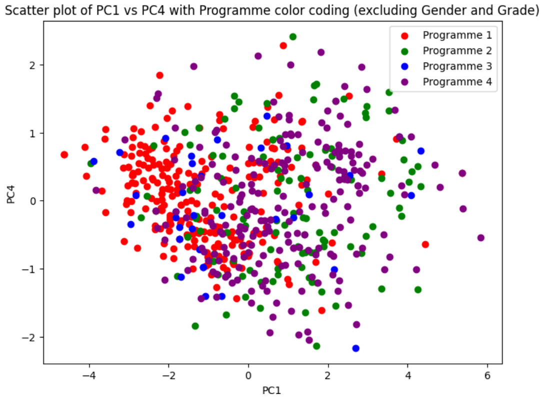

Fig. 5. scatter plot 1.1

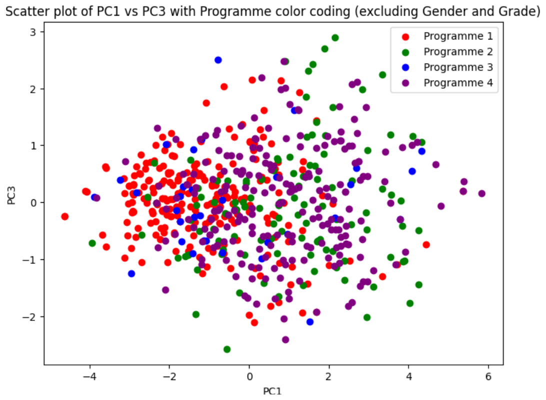

Fig. 6. scatter plot 1.2

This case is better since there is vague distinction between 3 to 4 kinds of o programs. However, such case does not successfully divide the four programs in different boundary regions in the box plot. Additionally, other sets of principle components under this circumstance are tried but the results do not have much differentiation while the best case is PC1 and PC3 which is shown in figure 6.

As a further step, various combinations of the column(always exclude index and program), as well as the pairs of principle components of them, are applied and generated the PCA scatter plot for observation and analysis. In the final analysis, the data chosen in the first experiment while including the column" grade" appear superior in performance.

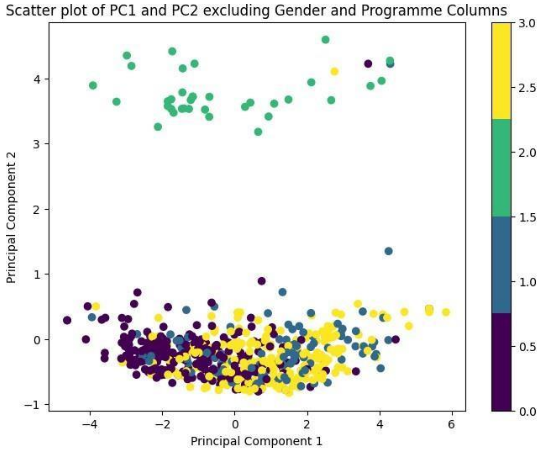

Fig. 7. scatter plot 2.0

Meanwhile, the same process is executed to the principal components to find out a more befitting case based on the box plot of principle components, hoping to find more ideal graphs. Finally, the PC1 and PC3 pair is found to be suitable.

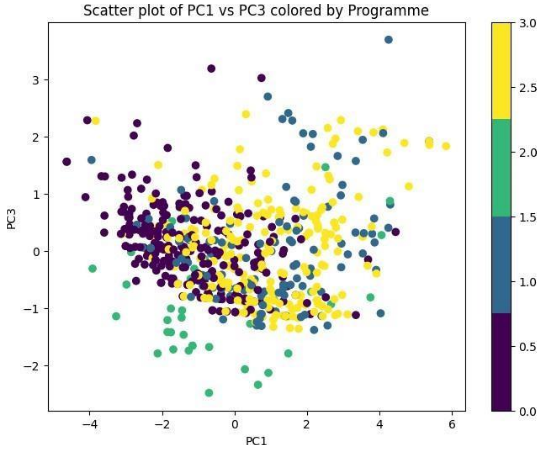

Fig. 8. scatter plot 2.1

III. Results

In conclusion, elements excluding gender and program are selected as best case to perform PCA. Under this circumstance, most of program features are basically classified. Program 1 and 4 have a tendency to increase on principle component 1. Specifically, program 3 are clearly separated from others.

On the other hand, one limitation is that program 1 is not yet divided during experiments. Moreover, apart from program 3, the remaining programs are merely visually tending to separate. There are some outliers, as well.

IV. Discussion

The prior method to achieve data dimension reduction used in this lab is Principal Component Analysis. The principal components it generates are linear combinations, which is not able to explain complex variables. The mapped data in the scatter plot probably cannot reach ideal condition. Driven by variance, PCA are possible to ignore significant information in the raw data.

In the following lab sessions, one recommendation is to apply classifiers to divide the data that have already separated and keep classify the rest. Furthermore, diverse approaches will be attempted for advanced classification.