Lab Report of Coursework2

Data Classification Methods

Hongkun Jiang, 2253854
TA: Xin Gao

***Abstract*—This is a lab report of INT104 Artificial Intelligence coursework 2. In this section, the sample data in coursework 1 will be classified via several methods of supervised learning, which include decision trees and random forests, SVM (Support Vector Machine), and Na¨ıve Bayes. Eventually, an ensemble classifier will be built up to improve the performance.**

***Index Terms*—decision tree and random forest, SVM, Na¨ıve Bayes, supervised learning, etc.**

I. Introduction

AFter the visual observation of data in coursework 1 by PCA in a scatter plot, more accurate classification is demanded to specifically group the features named supervised learning. The criteria to evaluate the research result are accuracy (keep 3 decimal places), F1 score as well as its variance (keep 2 decimal places).

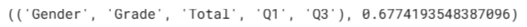
In the following section, which is the reach methodology, decision tree will firstly be applied. The advanced integration version of decision tree, which is random forest, will be tried after that to vote for result based on trees. Subsequently, SVM is adopted and various of feature combinations will be assessed. The comparisons of result data select the most suitable configuration. The next classification is Na¨ıve Bayes and reasons for feature set selection will be delivered. Meanwhile, ensemble classifier will mix the result of those classifiers and use for voting and optimization. Eventually, this report will state the final result of the experiment and do discussion.

II. Methodology

Within these classifiers, the accuracy is considered as the priority to select features and alter the parameters of different classifiers. However, the F1 scores, which is used to assess the balance of recall and accuracy, and its variance are also considered to ensure the extend of overfitting and bias.

A. Decision Tree and Random Forest

1) Decision Tree: To begin with, a decision tree classifier is built up and test the feature sets of data. For the first time, as y axis is set as program according to the purpose. The independent variable is the default feature sets in which index and program is excluded. It is not surprising that the accuracy is not ideal with the value of 0.609. Secondly, the experiment is trying to find the best feature set of columns based on the criterion of gini impurity. The parameters are set as: max depth=10, min samples split=20, min samples leaf=10 by default. These are not precise but basically make the model not to overfit it. Using iteration to test the combinations of the features, a final result is given as below:

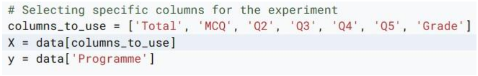


```
(('Gender', 'Grade', 'Total', 'Q1', 'Q3'), 0.6774193548387096)
```

Meanwhile, its f1 score and its corresponding variance, which are 0.67 and 0.06 can be described as effective. Therefore, this combination will be the final feature selected and the highest accuracy is 0.677.

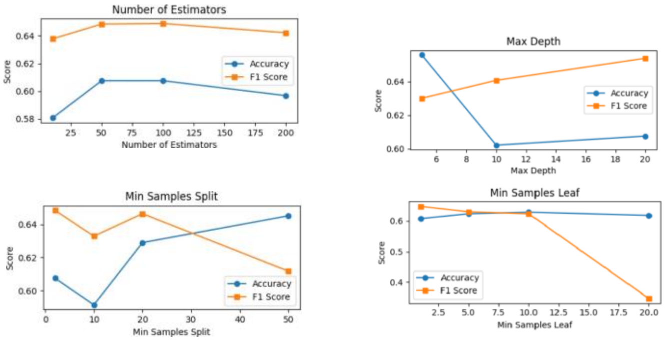
Nevertheless, a single decision tree is not able to represent a complex data file like this. It has still the risk of overfitting one case and instability even though the variance is low due to few experiments. To overcome, we adapt random forest to gain generalized result.

2) Random Forest: Firstly, the final feature set tried in decision trees is used. Surprisingly, this set do not make sense to random forest because the accuracy has merely reached 0.618 and F1 score is even lower as 0.57. Reason may be that the decision fits a special case and choose features not by dependency and random forest eliminate this. Therefore, according to the content of coursework 1, the feature that is more highly related to the program is selected. Additionally, since each subset of a tree in random forest tends to choose the most suitable features, by definition, more features should be chosen. Only index, gender and Q1 are dropped.

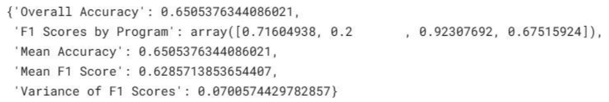


```
# Selecting specific columns for the experiment
columns_to_use = ['Total', 'MCQ', 'Q2', 'Q3', 'Q4', 'Q5', 'Grade']
X = data[columns_to_use]
y = data['Programme']
```

Within the parameters same as decision tree, the accuracy has increased to 0.640. The next step is fixed the parameters to gain higher accuracy. So here draws the line graph of accuracy and F1 score when estimators, depth, min sample spilt and min sample leaf changing.



According to the figure, the number of estimators (trees) is set as 50 because the level of accuracy and F1 score reaches top and 50 save computational cost. The accuracy decreases fast before 10 while F1 score increases smoothly. Concerning this and the overfitting issue, 5 is set as depth value. Min sample spilt is set to 30 for accuracy and F1 score balance, and for the same reason as depth min sample leaf is 10. Here is the result.



```
{'Overall Accuracy': 0.6505376344086021,
 'F1 Scores by Program': array([0.71604938, 0.2       , 0.92307692, 0.67515924]),
 'Mean Accuracy': 0.6505376344086021,
 'Mean F1 Score': 0.6285713853654407,
 'Variance of F1 Scores': 0.0700574429782857}
```

Consequently, the random forest model performs better if fixing the parameters accordingly and do not seem to overfitting with variance of 0.07. These data is not as ideal as random forest perhaps due to higher stability. And the clear classification of program 3 is corresponding with the PCA result.

B. Support Vector Machine

SVM works on a graph by calculating different distances of each pair of samples and decide the region of classification (soft margin) because the scatters cannot obviously divide. In this section, the PCA scatter plot of coursework 1 is reused since it is the best graph currently obtained, where SVM works on, that clearly classifies different programs.

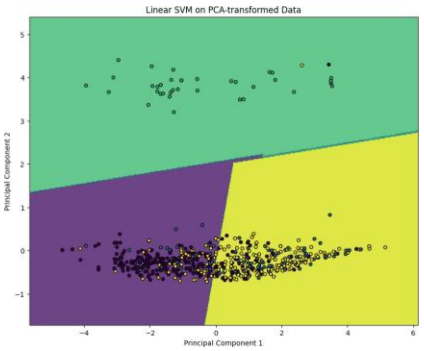


Fig. 1. Linear SVM

1) Linear SVM: The feature set is the same as random forest since it is from PCA result. The accuracy of linear SVM is 0.602 and F1 score is 0.56, which is lower than results in previous classifications but still acceptable. The strength of this classifier is variance close to 0 (0.0022 actually). Moreover, polynomial SVM is also attempted.

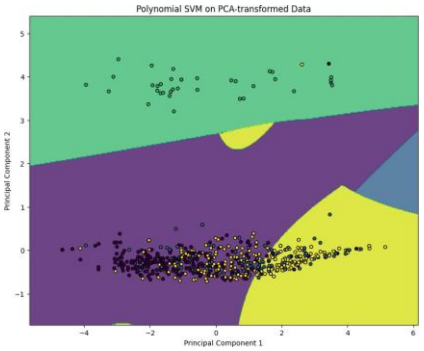


Fig. 2. polynomial SVM

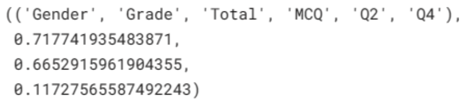
2) Polynomial SVM: The three resulted value, which are 0.601, 0.55 and 0.00, are strictly similar to it in linear SVM. It can be distinguished in the figures that the distribution of the 4 programs is regular. Program 3 are at the top and can be easily classified almost 100%. Program 1 and 4 below have a tendency in x axis but program 3 is mixed. This distribution matches the result in all classifications so far. SVM acts poorly in the case compared with others, but do not overfitting.

C. Na¨ıve Bayes

The Na¨ıve Bayes is a general function to classify. It assumes that all of the features in the data is independent and do not affect each other. Although this does not seem to be correct, Na¨ıve Bayes performs well in practice. There are specific classifiers of Na¨ıve Bayes.

1) Polynomial Na¨ıve Bayes: Concerning that the individual marks of questions are integers in a range and the basic information are discrete, the first consideration is polynomial Na¨ıve Bayes. Given that the fundamental principle of Na¨ıve Bayes and decision tree are both classifying groups based on the features and focus on the relevance, the same feature set from decision tree is selected. According to the experiment, the accuracy has only reached 0.540 and F1 score is even lower. When repeating the analysis by changing the feature sets, overall accuracy is found always lower than 0.6, which is an inferior outcome.

2) Gaussian Na¨ıve Bayes: The previous result has a trend of deteriorated performance metrics. The analysis indicates the possibility that the sample data may be better regarded as continuous and thus pointing to Gaussian Na¨ıve Bayes. This classifier holds an premise that each feature presents a normal distribution. Nonetheless, it may still perform well even if the condition unmatched. Likewise, feature set resulted from decision and PCA are used. The implement of the latter achieves improvement with the accuracy of 0.653. It represents promising potential to enhance the performance.



```
(('Gender', 'Grade', 'Total', 'MCQ', 'Q2', 'Q4'),
 0.717741935483871,
 0.6652915961904355,
 0.11727565587492243)
```

Fig. 3. Gaussian Na¨ıve Bayes

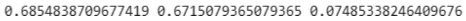
Iterating different combinations of the column, the experiment confirms outstanding results of 0.718 accurate rate. Though the variance is mildly higher, it is within acceptable limits.

D. Ensemble Classifier

As a further step, an ensemble classifier is supposed to be built. These sorts of classifiers are generalized due to the integration of its participants. Therefore, the major function is reduction of overfitting and somehow moderate the accuracy.

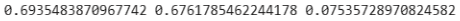
Initially, bagging, which uses voting to decide the result, is attempted. In light of the various type and character of different classifiers, soft voting is accepted as an attempt.

to increase accuracy. Soft voting means use probability as weighted result, while hard voting only cares about each final prediction. However, there are three general type of predictors so hard voting may has bias. Now, the most frequent dataset [’Grade’, ’MCQ’, ’Total’, ’Q2’, ’Q3’, ’Q4’, ’Q5’] of the 3 sub-classifiers is applied.



```
0.6854838709677419 0.6715079365079365 0.07485338246409676
```

Such output is exceled in performance, since it equips an ac- curacy close to the highest with satisfactory F1 score variance. By further experiment, here draws an optimized result.



```
0.6935483870967742 0.6761785462244178 0.07535728970824582
```

The feature set is slightly different without the feature” Total”. In summary, 0.694 is the highest value currently.

III. Conclusion

In conclusion, the classifiers tested in the lab achieved over 60 percent accuracy and acceptable f1 score as well as the variance. Random Forest performs standardized. SVM tend to dramatically reduce the F1 score variance to avoid ovfitting but acquires lower accuracy. On the contrary, Gaussian Na¨ıve Bayes, as the final result of Bayes classifier, shows excellent accuracy of 71% with the issue of little overfitting. Ensemble classifier of them enhance performance and keep balance of their strength. Eventually, the single decision tree is contingent and the best dataset of it is not reliable for other classifiers. Feature conbinations selected by the final result are highly coincident with their dependency with program column which implied in CW1, PCA. In addition, the experiment is restricted to supervised learning and the effect of cluster has not been proven yet.