Lab Report-CW2

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

May 8, 2024

1 Introduction

The purpose of this report is to verify the influence of different feature subsets of the same data set on the prediction results of the classifier model, and the influence of different model parameters on the predicted values of the same feature subset. Four main classifier models are used in the experiment, they are decision tree, random forest, svm and naive Bayes. Also included are two integrated classifier models, they are voting and Adaboost.

2 Methods

2.1 *decision tree*

First of all, the experimental model is the decision tree model. We divided 25% of the data set into the test set, and the default initial model training result was 48.38709%. Clearly this result is unsatisfactory, and the process of verifying the results is trained in an uneven manner. We then perform artificial model parameter adjustments based on the data set structure, which are {’criterion: entropy’, ’max depth’: 4, ’min samples leaf’: 12, ’min samples split’: 4}. The accuracy of parameter prediction of this model is 55.48387%. Although the prediction rate has improved significantly, the specific value is still not satisfactory. Therefore, we introduce the GridSearchCV object for ergodic comparison of model parameters and cross-network verification during model training. The parameters of the traversed model include the algorithm pattern, the maximum decision tree depth, the number of decision tree node samples and the number of internal node splits in the decision tree. After the traversal test of GridSearchCV object, the prediction accuracy can rise to 62.58064% for fixed feature subsets [’Grade’,’Total’, ’MCQ’,’Q1’,’Q2’,’Q3’,’Q4’,’Q5’]. The best combination of model parameters is {’criterion’: ’gini’, ’max depth’: 2, ’min samples leaf’: 1, ’min samples split’: 2}. In addition, a 3-fold cross-validation of the model training stage is carried out to make the prediction rate more accurate. The confusion matrix of the prediction results of the three models is compared **(figure 1)** .

We can see that the prediction results of different decision trees are different after parameter adjustment, and the number of correct data classifications for GridSearchCV object operation is relatively increased. Looking at the parameters of the three decision trees, we can see that a deeper tree may overfit the training data, while a shallower tree may underfit. As mentioned above, the depth of the second tree is too large, resulting in overfitting training, while the depth of the third tree is relatively shallow, so it has higher prediction accuracy.

2.2 *random forest*

Next, we use a higher-level application of decision trees, random forests, to continue our experiment. This time we mainly focus on the influence of different feature subsets on the prediction accuracy. Again, we use the GridSearchCV object to help us select model parameters and cross-network validation. The

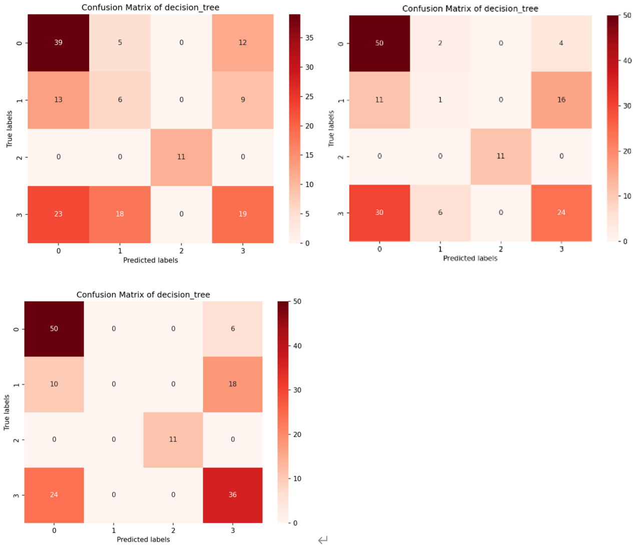

Figure 1. The upper left is the default model parameters, the upper right is the manually adjusted model, and the lower left is the model after the GridSearchCV object operation

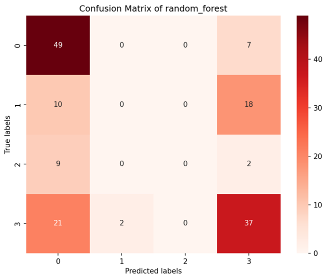

Figure 2. The confusion matrix after ’Grade’ was removed

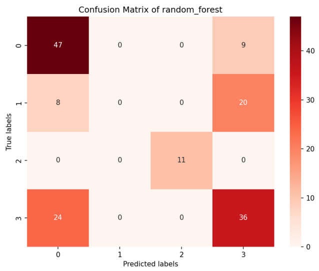

Figure 3. The final complete model is trained out of the confusion matrix

best verified parameter set for this fixed subset of features is ’max depth’: 3, ’min samples leaf’: 3, ’min samples split’: 13, ’n estimators’: 18. When selecting different feature subsets, Index and Gender are first excluded from the feature subsets in the data preprocessing stage. To the remaining subset A: [’ Grade ’, ’Total’, ’MCQ’, ’Q1’, ’Q2’, ’Q3’, ’Q4’, ’Q5] training data model, it is concluded that the prediction accuracy is 59.91935%. Then, by using the single variable method, we delete each feature in subset A, and input the deleted feature subset into the model. The prediction accuracy is 52.90322%, 63.64516%, 61.87096%, 60.13548%, 61.93548%, 60.64516%, 63.87096%, 61.29032%. Through observation and analysis, we can find that the prediction accuracy is significantly improved after the deletion of some features, such as ’Total’, ’MCQ’,’Q2’ and ’Q4’, while the prediction accuracy is greatly decreased after the deletion of some features, such as ’Grade’. This is very confusing, and by observing the confusion matrix, we can see that ’Grade ’features have a significant impact on the predictions of programme 3. Below is the confusion matrix **(figure 2)** with ’Grade’ removed.

At the same time, those features that have little impact on the prediction accuracy, such as’ Q1 ’, ’Q3’ and ’Q5’, were deleted to avoid interference with the prediction. In summary, the feature set [’Grade’,’Total’, ’MCQ’,’Q2’,’Q4’] is the most suitable prediction feature set, with a prediction accuracy of 65.29032%, and its confusion matrix is shown in the figure**(figure 3)**.

2.3 SVM

The same approach is applied to the SVM model. GridSearchCV object is used for traversal adjustment of model parameters and network cross-validation. The final adjustment parameter is {’C’: 1, ’coef0’: 0.9, ’gamma’: ’scale’, ’kernel’: ’rbf’}, where the main c value is required. Because it affects the degree of model regularization, when C is large, the model will pay more attention to reducing the number of misclassified samples, which may lead to better performance of the model on training data, but the generalization ability may be poor, and it is easy to overfit. When C is small, the model will pay more attention to choosing hyperplanes with larger boundaries, which may lead to higher tolerance for misclassified samples and better generalization ability, but the performance on training data may not be as good as when C is large. After the traversal range control of the model parameters, the prediction accuracy is the highest when the c value is [0.5, 1.5] for the same feature subset. After many tests, the experimental feature set [’Grade’,’Total’, ’MCQ’,’Q1’,’Q2’,’Q4’] can get the highest prediction accuracy of 65.80645% **(figure 4)** .

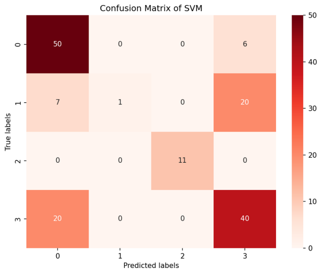

Figure 4. The confusion matrix of the final SVM model

2.4 Naive bayes

In the naive Bayes model, we chose a Gaussian kernel. The characteristic of this model is that it runs fast compared to other models. For the feature subset selection of this model, we adopt a way to preserve the features that have a greater impact on the target variable. The L1 regularization of Lasso method makes the coefficients of some features become zero, thus achieving sparsity. For feature subset [’ Grade ’, ’Total’, ’MCQ’, ’Q1’, ’Q2’, ’Q3’, ’Q4’, ’Q5] in the regularization parameter of 0.05, feature selection is the result of the Lasso model [’ MCQ’, ’Q2’, ’Q4’] these characteristics has a higher effect on forecasting. However, when it was finally sent into the model training, it was found that the prediction accuracy of the subset had a certain deviation, only 55.29839%. Obviously, this is unacceptable. By observing its confusion matrix, it can be concluded that there is a significant deviation in the prediction accuracy of programme 3. By observing the composition of the data set, we found that due to the particularity of grade characteristics, its relatively concentrated data distribution could not reflect the overall distribution in Lasso model. However, for programme 3, there is a fixed grade feature value corresponding to its label, so for the overall model classification, grade should be added to the feature subset. At the same time, other possible feature subsets are tested, and it is found that for naive Bayes. Therefore, for the feature subsets [’Grade’, ’MCQ’,’Q2’,’Q4’], the final prediction accuracy of the naive Bayes model reaches 64.22580%, and the confusion matrix is shown below **(figure 5)**.

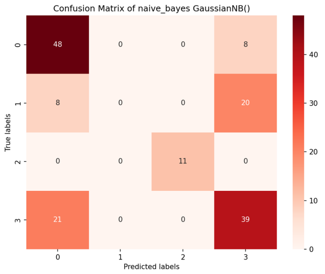

Figure 5. The confusion matrix of the final Naive bayes model

2.5 Ensemble classifier

For the integrated classifier, two methods were used in the experiment to conduct comparative experiments, namely voting and adaboost. The first is the voting model, whose weak learner we chose the four classifiers from the previous experiment, decision tree, random forest, svm, and naive Bayes. In order to improve the accuracy of the predicted value, the parameters of the weak learner are modified to the parameters corresponding to the best prediction rate of the previously trained model. After multiple combination tests, the feature subset of the best accuracy was determined as soft voting for [’Grade’,’Total’, ’MCQ’,’Q2’, ’Q4’,’Q5’]. The final prediction result was based on the weighted average of the prediction probability of each model, and the accuracy of the training result was 63.22580%. The confusion matrix is shown in **figure 6**.
Then there is the adaboost model. The AdaBoost algorithm can focus on those samples that have been misclassified before so that the subsequent basic model (weak classifier) will pay more attention to these samples that are difficult to classify so that the misclassified information will reduce the impact on the subsequent

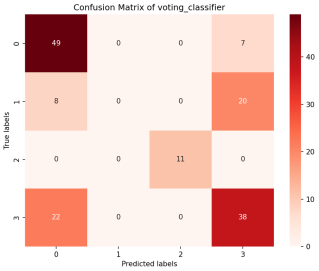

Figure 6. voting model confusion matrix

classification. At the same time, we set the learning rate to 0.75 so that the weights of the previous round of weak classifiers are scaled in each iteration to control their contribution to the final integrated model. After many combination tests, we found that whether the feature subsets [’Grade’,’Total’, ’Q2’,’Q3’,’Q4’,’Q5’] contain the feature ’MCQ ’,’ Q2’,’ Q4’ does not have a big impact on the prediction rate. This is different from previous conclusions about the effect of corresponding characteristics. We hypothesize that the lifting algorithm makes up for that. The final prediction accuracy is 68.58064%, and the confusion matrix is shown in the figure below **(figure 7)**.

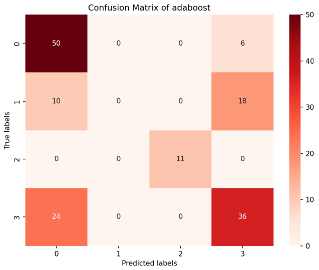

Figure 7. adaboost model confusion matrix

3 Conclusion

In this study, decision trees, random forests, SVM, naive Bayes and integrated classification algorithms were used to predict the data. GridSearchCV object is used for ergodic comparison of model parameters and network cross-validation, and the best combination of model parameters is found. At the same time, the feature subset which has the greatest influence on the prediction accuracy is determined by feature selection and the single variable method. In summary, feature subsets containing features ’Grade’,’Total’, ’MCQ’,’Q2’,’Q4’ can improve the prediction accuracy of the classifier.