COMP2261 Artificial Intelligence–Machine Learning 2021/2022 (revised September 2021)

Module Name: <COMP2261>
Date: <23rd January 2022>
Submitted as part of the degree of BSc Computer Science
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Describe the approach for solving the supervised learning problem by predicting gender using three classification machine learning algorithms. Male and female genders are predicted. It is envisaged that the solution to the challenge will assist individuals in predicting gender based on certain humour style scores and question scores. The algorithms used are Random Forest Classification, Support Vector Machine (SVM) and Logistic Regression. Support Vector Machine generally shows better performance than the other algorithms in this problem.

Index Terms—Supervised Machine learning,sklearn,classification,Random Forest Classification, Support Vector Machine, Logistic Regression, gender, humour style

1 Introduction

THERE has been a dramatic increase in the number of teenagers who are unsure of their gender, whether they identify as female, male, non-binary, or any of the other numerous categories used on the gender spectrum in recent years[3]. Meanwhile, the NHS provides assistance with gender testing[3]. A previous study established a link between the humour style score and gender[2]. If we can use some questions and humor styles to predict gender, it can be used in certain instances to help people know what their gender is. This report uses data about humour style scores as features to predict gender by applying three classification algorithms. This would help people understand their gender better.

2 Related Work

The dataset by Martin et al. in 2003 consists of 1071 rows and 39 columns, containing 32 questions, humour style scores, age, gender and accuracy[2]. These questions can be used to determine four distinct humour styles. Each question has a minimum score of -1, which means the person did not response the question, and a maximum score of 5. Higher score means more often and vice versa. Additionally, gender has three values: 1 for men, 2 for females, 3 for others, and 0 for no response. There are four types of humour mentioned: positive use of humour to reinforce oneself (Self-enhancing), aggressive use of humour to reinforce oneself while disparaging others (Aggressive), positive use of humour to strengthen relationships while demean oneself (Self-defeating), and positive use of humour to enhance one's relationship with others (Affiliative)[2].

3 METHODOLOGY

3.1 Data Analysis

According to figure 1, there are 581 males, 477 females, 8 others, and 5 no response. These results are expressed as a percentage of the total data as follows: 54%, 44%, 0.75%, and 0.47%, respectively. This indicates that the majority of the population is composed of males and females. The lowest response to age is 14 and the highest is 44849, which is clearly unreliable . The lowest accuracy is 2% and the highest is 1, although it is noteworthy that the average is 87.5% and there are only 16 cases where the accuracy is less than 50%. And only 658 rows have a greater accuracy than the average. The majority of accuracy responses fall between 70% and 100%, which corresponds to 979 pieces of data and 91% of the total data. In general, all humour scores fell fairly close to the mean, which is 3, but the exception of affiliative is 4.

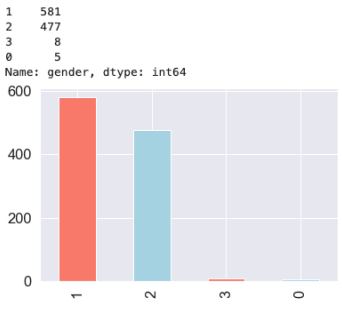

Fig. 1. Gender statistics

3.2 Pre-processing Data

Firstly, remove all rows which had a gender response of 0 or 3 because there are not enough relevant data. Secondly, remove data for all ages above 100 as unreliable reason. Thirdly, remove rows for all accuracy scores below than 0.7, this is done without losing data and without reducing the accuracy of module. Next, selecting features by applying 'SelectFromModel' function. This function is created by filtering the coefficients in order to acquire the features, and the coefficients are obtained by running an algorithm that returns each feature coefficient for comparison. In this time, the features we require have coefficients greater than the average. The features that were selected according to Logistic Regression and the data were normalised by using Standard Scaler. After selecting the features, check the balance of features according to the histogram, remove features which are clearly unbalanced. Finally, Q11 , Q12 , Q15 , Q19 , Q23 , Q3 , Q31 , Q4, agressive and selfenhancing are retained as features.

3.3 Post-processing observations

Firstly, according to the histogram from figure 2 and plot from figure 3, it is clear to see that all the features are almost balanced, suggesting that it is able to apply Standard Scaler after. Secondly, the predicted sample size for males is reduced to 513, or 55% of the total, and the predicted sample size for females is reduced to 415, or 45% of the total. The ratio of the number of males to females is 1.2, indicating that the target data is approximately balanced. According to the correlation matrix in Figure 4, it can be seen that Q12, Q19, Q3, Q14 and aggression have the strongest correlation scores with gender, at 0.148, 0.261, 0.135, 0.113 and 0.145 respectively; for other features(Q11, Q15, Q23, Q31), they do not score too badly, at 0.070, -0.040, -0.109 and - 0.004. There are also features that have strong correlation score with other features as well, for example Q15 has 0.426 with Q23, indicating that there are some similarity features. After cleaning the data, the size of the dataset has been reduced to 928 rows and 5 columns.

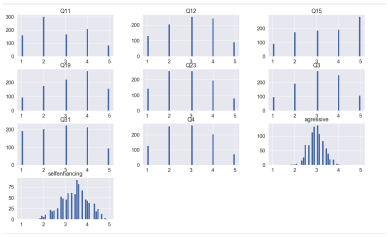

Fig. 2. Features Histogram

3.4 Data splitting and scaling

Splitting the data into a training set and a test set, with the test set taking up 20 percent of the data. As discussed in Exploratory Data Analysis section, nearly every feature satisfies normal distribution. So applying Standard Scaler will be the better way, which is good for the normal distribution[1]. From the figure 3, by comparing the data before and after the scaler, it is evident that the data now has a more normal distribution.

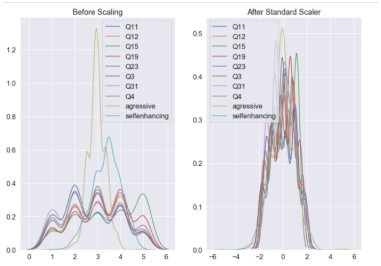

Fig. 3. Plot about Before and after scaler

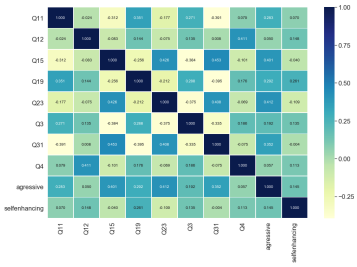

Fig. 4. Correlation matrix

3.5 Algorithm Selection

The first classification algorithm is Support Vector Machine(SVM), which is really suitable to classifying the complex but small or medium dataset[1]. It is also really sensitive to the feature scales, which means after scaling, the result normally will be much better[1]. For this dataset, as discussed before, there are lots of similar features, and the size of dataset is quite small(size only 928). As a result, the Support Vector Machine was chosen.

The second classification algorithm is Random Forest Classification, an algorithm involves an ensemble of decision trees and employs bagging, which takes samples with replacement[1]. Then passed the result to Decision Tree Classifier to get better performance[1]. The reason to choose Random Forest Classification is because even though the data is missing or inaccurate, Random Forest Classification still can maintain its accuracy[1]. For this dataset, although the accuracy rate is clear, which part of the data is biased still unknown. Applying this algorithm may compensate for this, so choose Random Forest Classification as the superior option.

The final classification algorithm is Logistic Regression, a regression algorithm but also can be applied in classification problem[1]. It normally makes a binary classifier[1]. The algorithm computes a weighted sum of the input features and return logistic as result[1]. For this data, since only male or female is predicted, it is like a binary classifier (male or non-male). So this algorithm is suitable in this situation.

4 Evaluation

4.1 Evaluation metrics

4.1.1 Accuracy

The model accuracy is the basic metric to measure the correct prediction of the algorithm, which is the number of

correct predictions made as a ratio of all predictions made. The formula is following[1]:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total numbers of predictions}} \quad (1)$$

Normally, the higher accuracy means the better module performance. However, it is often misused[1]. Accuracy only can be directly used as there are an equal number of observations in each class[1]. To be able to get better evaluation, cross validation can be a better way to do further improvement.

4.1.2 Confusion matrix

The confusion matrix returns true negatives, false positives, false negatives and true positives. A good module should have more values in true negatives and false negatives[1]. By using these four values, there are two more metrics can be used, which are precision and recall(sensitivity). Following are the formula of these:

$$\text{precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{recall(sensitivity)} = \frac{TP}{TP + FN} \quad (3)$$

Specifically, TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives.
Usually, these two metrics need to be used in conjunction. Higher precision also means lower recall and vice versa[1]. In this problem, higher recall is important, because it is necessary to provide people with an efficient way of identifying gender, and people do not want to spend a lot of time getting an ambiguous answer. However, higher precision also cannot be ignored, since people are often troubled by the fact that gender is often a minuscule difference, so the approach needs to possess specificity. In this situation, it is necessary to find balance between these two metrics.
By combining these two scores, then there is another score that comes out, which is f1 score. The F1 score is the harmonic mean of precision and recall. Following is the formula[1]:

$$\text{f1} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{FN+FP}{2}} \quad (4)$$

Higher The f1 score means that have more similar precision and recall[1]. In this case, better f1 score could means better performance.

4.1.3 K-fold Cross Validation

K-fold Cross Validation is a technique that splitting the training set into K-folds and making predictions for each fold[1]. The score can be varied; for instance, in a classification problem, the score can be accuracy, f1, precision, or recall. The reason for using Cross Validation is because it provides the same performance as adding a validation set to verify the output but is considerably more convenient[1].

4.1.4 Receiver Operating Characteristic(ROC) Curve

The Receiver Operating Characteristic (ROC) curve plots sensitivity (recall) versus specificity. The area beneath the curve represents the module's performance[1]. The area has a value between 0 and 1. Generally, the region closest to 1 indicates more accuracy[1]. Therefore, in future evaluations, it is beneficial and straightforward to compare the size of the under area to determine the module performance.

4.2 Model Training and Evaluation

4.2.1 Support Vector Machine

The parameters used are the margin and hyperplane. C and gamma are the primary hyperparameters. C is in control of the breadth of the selected data and the number of margin violations. For instance, a lower C value indicates a bigger data selection area but more margin violations, whereas a higher C value indicates fewer margin violations[1]. Thus, modifying C is an efficient method to avoid over- and under-fitting[1]. For example, by lowering C in order to avoid overfitting[1]. Gamma is used to adjust the sphere of influence; for example, increasing the gamma to narrow the impact area[1]. As with C, gamma can be a useful tool for avoiding over- and under-fitting. For instance, by boosting gamma in order to prevent underfitting[1]. To tune the hyperparameters, select 30 random numbers from 0 to 1 as the value of C and gamma to ['scale','auto'], and then do a grid search and 10-fold cross-validation to determine the optimal hyperparameters. Cross-validation is used instead of conventional accuracy to avoid overfitting. The figures 5 illustrate the matrices obtained from the confusion matrix. As can be noticed, it has a high recall score and a no much difference precision score, indicating that they have a good balance with it. Additionally, the f1 score is greater than the accuracy, indicating that the module is not over- or under-fitted. Applying the ROC curve reveals that the area is 0.66, indicating that the accuracy is acceptable in this circumstance.

	model	cv_acc	cv_precision	cv_recall	cv_f1
0	gs_lg	0.610963	0.633455	0.703846	0.666253
1	gs_rfc	0.609935	0.623561	0.723115	0.666330
2	gs_svc	0.614236	0.619875	0.781448	0.690739

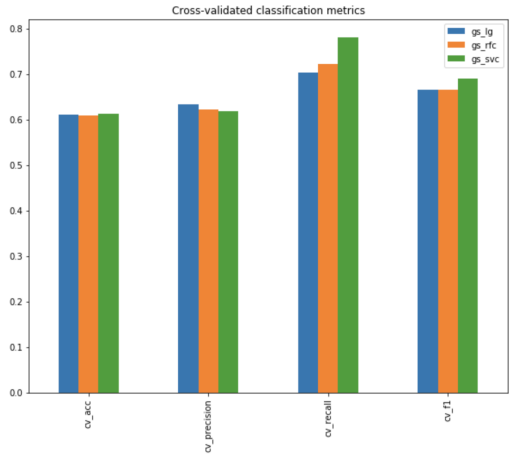

Fig. 5. Matrix result

4.2.2 Logistic Regression

The parameters used are the weight of each feature and the coefficients. The key hyperparameter is C, which specifies the degree of regularisation; the lower the degree of regularisation, the greater the value of C[1]. It must be positive, and C is set to 1.0 by default. As with the Support Vector Classifier, it may prevent over- or under-fitting by modifying the C value. To tune the hyperparameter, select 30 random numbers from 0 to 1 as the value of C and repeat the SVC tuning procedure. The optimal value of C is 0.62. The figures 5 represent the matrices obtained from the confusion matrix. Both accuracy and recall get a high score with a little difference, indicating that they are in excellent balance. Additionally, the f1 score exceeds the accuracy, showing that the module is not over- or under-fit. The area of the ROC curve is 0.67, showing that the accuracy is adequate in this case.

4.2.3 Random forest Classification

This module is a nonparametric model, which means that the number of parameters is not defined before to training, and we often set the usage parameter to limit the Decision Tree's flexibility in order to minimise over-fitting[1]. The parameters are the tree's breadth and depth, as well as the number of nodes. Max-depth and n-estimators are the main hyperparameters. Because max-depth controls the maximum depth of regularisation, it may be utilised to limit the risk of overfitting by lowering max-depth. n-estimators provide more control over the number of decision trees, which improves accuracy[1]. It also regulates the rate at which the process is done, thus a compromise between precision and speed must be established. Set max-depth to [5,10,8,9,15], n-estimator to [300,250,150,280,290], and repeat the tuning procedure as described previously. The figures 5 demonstrate the matrix obtained from the confusion matrix. Precision and recall both have high ratings with limited variance, suggesting that they are in balance. Additionally, the f1 score is greater than the accuracy, indicating that the module is neither overfit nor underfit. The ROC curve indicates that the area is 0.64, indicating that the accuracy is acceptable in this situation.

4.3 Model Comparison

Overall, the findings for the three modules are quite promising, despite the low accuracy(above 60%). In cross-validation and testing, the models utilised attained a high and balanced sensitivity(above 70%) and precision(higher than accuracy).

As seen in the figure 5, Random Forest Classification has a better degree of accuracy and precision. It is worth mentioning, however, that the variation in accuracy between the three is not significant. The difference is entirely between 1% and 2%. In other words, the accuracy and precision of all three models are fairly similar. Additionally, Support Vector Machine has the greatest recall (Sensitivity) score, with a difference of 8% between the highest and lowest. implying that this module performs better in unusual conditions. It is worth mentioning, however, that this module has the lowest precision, with a difference of 2% from the other modules. Additionally, the module has the highest f1 score, which is 3% higher than the others. This indicates that the module possesses a healthy balance of sensitivity and precision. In general, Support Vector Machine outperforms the others due to its excellent balance of accuracy and recall, as well as its relatively high accuracy when compared to other models.

5 Conclusion

The purpose of this report is to make gender predictions using scores on a variety of questions and selected humour style scores. Due to a lack of data, the estimated gender will be either male or female. The report began by visualising the data and then removing unnecessary data. The report then analysed the data to ascertain its balance and discussed scaling the data using the Standard Scaler. Additionally, the study justified the selection of three algorithms and few matrices. Following that, the study discussed the training and turning processes in detail, including an explanation of hyperparameters and matrix scores. At the conclusion of the report, it reviewed the performance of three algorithms on various matrices and concluded that Support Vector Machine performs the best. However, this procedure is not without limitations. Firstly, there are no more gender categories, which implies that the module cannot anticipate individuals who are transgender, for example. Secondly, while cleaning up the data, we deleted any questions with a score of -1, which resulted in the loss of some data. Thirdly, it requires certain humour style scores, which might be a bit complicated to obtain immediately. Fourthly, Only one algorithm was used to choose features, which may not be relevant to the others. The following improvements can be made to the work in the future. Firstly, collecting data on people with different gender types. Secondly, By substituting gender-related questions for humour style scores, precision and convenience may be improved. Thirdly, cleaning data could try to use mean value replace the missing value. Fourthly, experiment with alternative methods for selecting characteristics and prioritise the most common ones. After this project, I have better understanding on the process of machine learning, including data analysis, algorithm selection and algorithm evaluation. In future, these modules aim to help people have better understand about gender.

References

[1] Geron, A. Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. (O'reilly Media,2017)

[2] Martin, R., Puhlik-Doris, P., Larsen, G., Gray, J. & Weir, K. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal Of Research In Personality*. **37** pp. 48-75 (2003,2)

[3] NHS Worried about your gender identity? Advice for teenagers. (nhs.uk,2018,6), https://www.nhs.uk/live-well/healthy-body/trans-teenager/