Prediction of Wine Quality Using Ensemble Learning Approach of Machine Learning

Qingwen Zeng

Institute of Disaster Prevention, department of Information Engineering SanHe city, China, 06500
[*]Corresponding author. Email: wuqiusuozqw@163.com

Abstract

The improvement of consumption level leads to increase in demand for red wine.What follows with is the contradiction between production speed and quality control. Predicting red wine quality in traditional process demands a lot of time and labor costs makes the whole productive process more expensive. Nowadays, benefit from Machine Learning (ML), especially the rise of ensemble learning, red wine quality prediction could have a more efficient and more convenient way. During this process, use a certain amount of data of several specific features to be trained by ensemble learning model to find the best result could be used in prediction of the red wine quality. The best model combination we found is stacking ensemble learning with accuracy rate of 0.87. This research could be a significant reference for red wine test or further use in the related industrial manufacture to reduce the cost of quality production.

Keywords-*wine quality; data analysis; machine learning; ensemble learning*

1.INTRODUCTION

Red wine as one of the most famous liquor in the world, the economic benefits that red wine bring are enormous. The traditional way to predict red wine quality includes three parts:sight, smell and taste. All of them need to be certificated by people with years of professional training which already cost many resource, time and money, not to mention the wine quality test only can be accomplished after whole production process ended. What industrial production need is a technology that can perform quality identification at any time. The quality of red wine may cannot be identified directly in production process, but the content of each component of red wine can be detected as a data to predict the quality of red wine [1], which is exactly in line with the idea of machine learning.

ML can be appropriately applied to most aspects of modern production, machine learning technology is mature enough to make it happen. Before this paper, there were some studies related to predict wine quality by machine learning [2-4], they used a variety of machine learning models to achieve very good results, but there is still possibility for improvement in accuracy. Although machine learning models are very powerful, but a single model always has limitations. Ensemble learning as an algorithm that can fuse multiple models that has performed well in the field of ML provide a way to breakthrough these limitations to get higher accuracy rate.

This paper is devoted to exploring how to improve the accuracy of wine quality prediction by ensemble learning. Higher accuracy can better serve as a reference for red wine quality prediction, which provides an idea for how to detect red wine quality during production, and reduce production losses.

The paper is organized as follows: In Section 3, we perform data processing, includes discussion of datasets, data analysis and feature engineering. Section 4 introduces ensemble learning and the structure of ensemble learning used in this paper. Section 5 fits the model to the data and compare the gap between a individual model and an ensemble learning model. Section 6 summarizes the main finding and conclusion.

2.METHODOLOGY

2.1. Dataset and the Analysis

In this work, the red wine data used to be a large dataset containing red and white wine data on UCL Machine Learning Repository, but which the red wine data was republished on Kaggle which is public and

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 770–774, 2023.
https://doi.org/10.2991/978-94-6463-042-8_110

widely used. The red wine dataset contains 1599 samples, each sample consists of 11 physiochemical properties: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates, alcohol. In addition to these 11 properties, Quality as the output variable has been divided into 11 levels from 0 (bad) to 10 (excellent) by sensory data from several different professional testers [5].

Through data analysis, it is more clearly to describe the data characteristics of each variable in order to better understand how to process the data. The basic statistical analysis of variables is presented in the Table 1.

TABLE 1. STATISTICAL ANALYSIS OF VARIABLES IN THE RED WINE DATA.

Variable name	Minimum	Maximum	Mean	Std. Deviation	Median
Fixed acidity	4.600	15.90	8.320	1.741	7.900
Volatile acidity	0.120	1.580	0.528	0.179	0.520
Citric acid	0.000	1.000	0.271	0.195	0.260
Residual sugar	0.900	15.50	2.539	1.410	2.200
Chlorides	0.012	0.611	0.087	0.047	0.790
Freesulfur dioxide	1.000	72.00	15.87	10.46	14.00
Total sulfurdioxide	6.000	289.0	46.46	32.89	38.00
Density	0.990	1.004	0.996	0.002	0.997
pH	2.740	4.010	3.311	0.154	3.310
Sulphates	0.330	2.000	0.658	0.169	0.620
Alcohol	8.400	14.90	10.42	1.066	10.20
Quality	3.000	8.000	5.640	0.808	6.000

2.2. Feature Engineering

Feature engineering is the process of transforming raw features into features that better express the essence of the problem to improve the accuracy of model predictions. According to data analysis in Table 1. The range of values for different variables varies greatly. For example, the maximum and minimum values of citric acid is 1 and 0, however the range of Fixed acidity from 4.6 to 15.9. ML models are very sensitive, such a situation can cause the influence of each variable to ML model to be biased towards the side with a larger range which will affecting the accuracy of the model. Therefore feature scaling will be a necessary step. Meanwhile feature selection also affects the accuracy of the model to a large extent. This paper chose these two techniques for feature engineering.

The Pearson correlation coefficient (r) is a reference indicator that can help understand the relationship between features and response variables which is used to perform feature selection, it measures the linear correlation between variables [6]. The Pearson correlation coefficient (r) of the feature with quality shows in Table 2. Except Pearson correlation coefficient there are many complex methods to analyze relationship between features, which can be referred in paper "Selection of important features and predicting wine quality using machine learning techniques" [7].

TABLE 2. THE PEARSON CORRELATION COEFFICIENT (R) OF THE FEATURE WITH QUALITY

Fixed acidity	0.124	chlorides	-0.129	pH	-0.058
Volatile acidity	-0.391	Freesulfur dioxide	-0.051	Sulphates	0.251
Citric acid	0.226	Total sulfurdioxide	-0.185	Alcohol	0.476
Residual sugar	0.140	Density	-0.175		

To solve the huge gap between different feature values, the most common approach is feature scaling a method to compress all feature values to the same range which could be implemented by 0-1 normalization. 0-1 normalization is to scale the feature values between 0 and 1 [8]. Through this methods the weights between features will be more average to improve the accuracy of the model.

Except feature engineering smoothing [9] also be used to reduce the impact of noise, amplify the effect of important data.

3.ENSEMBLE LEARNING

In ML models, we always want a stable model that robust in every aspect, but single model is often not so high-powered, we can only get multiple models with preference respectively. Ensemble learning is to combine multiple models in order to obtain a model with a better predictive performance than any individual model. Ensemble learning can complement the

advantages and disadvantages of different models, usually has a higher accuracy rate, stability and less prone to the overfitting. The commonly used ensemble methods are bagging, boosting and stacking [10].

Bagging base on the bootstrap. In Bagging, bootstrap is used to take the replacement sampling from the overall dataset to get N datasets, and base on the dataset builds a model, the final prediction result is obtained by using the output of the N models, usually: classification problem uses the voting of N models to predict, and the regression problem uses the average of N models to predict.

Boosting also divide N datasets, in the beginning the same weight will be set to each datasets, then use the algorithm to train the training set for t rounds. After each training, assign a larger weight to the training examples that fail, lead the learning algorithm pay more attention to the wrong samples after each learning, finally obtain multiple prediction function.

Stacking usually be built by different models,obtain a ultimate predictions through combine predictions from several other learning algorithms. It's the main ensemble method be used in this paper.

The stacking contains one or more base models and a meta-model, dataset will be divided into N datasets, base models will fit data from the N datasets respectively and generate the predictions, then meta-model learn how to best incorporate models that predict outcomes from base models. Base models are considered to be level-0 models, while meta-model are considered to be level-1 model.

After multiple experiments, we have selected the following models to construct the stacking model in this paper: Logistic Regression, MLPClassifier [11], XGBClassifier [12], Random Forest [13], SVM [14]. All models are referenced from packages of the Sklean library. MLPClassifier, XGBClassifier, RandomForest are combined as base models, LogisticRegression as meta-model. Each model has it's own parameters be adjusted through the Grid SearchCV a frequently used and efficient parameter adjusting method. The value of N is 5. Specific ensemble structure chart is showed in Figure 1.

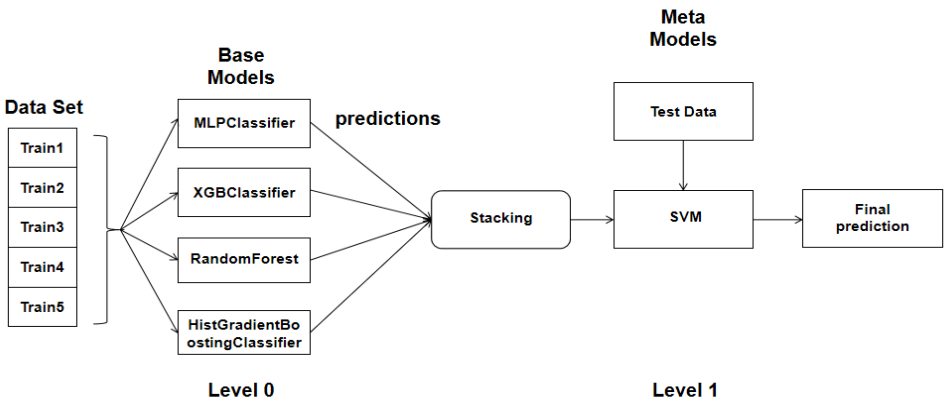

Figure 1. The structure chart of stacking ensemble learning using in this paper

4.Results

The model performance in machine learning classification problems is always measured using the following performance evaluation: accuracy, precision, recall, F1 score. They are defined as below:

$$\text{accuracy} = \frac{TP+TN}{TP+TN\_FP\_FN} \quad (1)$$

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

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

$$\text{F1 score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \quad (4)$$

In order to highlight the advantages of stacking, the performance evaluation of the model as a component is also calculated for comparison. The results are presented in Table 3.

TABLE 3. PERFORMANCE EVALUATION OF EACH MODEL

	Stacking	XGBClassifier	RandomForest	SVM	MLPClassifier
Accuracy(train)	0.999269	1.000000	0.997807	0.735842	0.750091
Accuracy(test)	0.870274	0.862120	0.853965	0.696071	0.712395
Precision(train)	0.999244	1.000000	0.997818	0.716683	0.734392
Precision(test)	0.872287	0.861250	0.851263	0.682735	0.703291

Recall(train)	0.999255	1.000000	0.997576	0.731786	0.746231
Recall(test)	0.875592	0.868280	0.860341	0.704371	0.720787
F1 score(train)	0.999269	1.000000	0.997806	0.719675	0.741520
F1 score(test)	0.868255	0.857931	0.848361	0.677635	0.700606

As presented in Table 3, ensemble learning Models (XGBClassifier, RandomForest, Stacking) have a 10%~15% rate advantage over a single model (MLPClassifier, SVM) in every performance evaluation. The model with the highest performance is stacking, XGBClassifeir followed by Stacking about 1%, however XGBClassifeir performed too well on the train set with 100% in every performance evaluation that means it has severe overfitting problem. Although there is no overfitting problem in RandomForest, there is still a gap of about 2% with stacking in terms of accuracy. Stacking combines the advantages of these models not only with the highest accuracy but also avoid overfitting problems.

5.CONCLUSION

This work uses ensemble learning approach to predict red wine quality. Base on the same data, stacking shows very powerful performance: the highest performance, the less overfitting. Nowdays even a 1% improvement could save the cost of mass production is incalculable. It is difficult for us to rule out the negative effects of these small datasets, but ensemble learning is undoubtedly a good choice for making predictions on red wine datasets. If we switch to a better dataset or do data feature engineering more thoroughly may could get a better effect which is future directions of this work. Through this work people can better predict the quality of red wine before the wine

making process ends, be able to give some objective reference standards for red wine production.

ACKNOWLEDGEMENTS

I have received many guidance and assistance from Prof. Pietro Lio, who I would particularly like to acknowledge. With his observant tuition and advice, this paper was completed successfully.

REFERENCES

[1] H. Song, J. Choi, C. W. Park, et al. Study of quality control of traditional wine using it sensing technology [J] Journal of the Korean Society of Food Science and Nutrition, 2015, 44(6): 904-911.

[2] B. Shaw, A. K. Suman, B. Chakraborty, Wine quality analysis using machine learning [M] Emerging technology in modelling and graphics. Springer, Singapore, 2020, pp.239-247.

[3] K. R. Dahal, J. N. Dahal, H. Banjade, et al. Prediction of wine quality using machine learning algorithms [J] Open Journal of Statistics, 2021, 11(2): 278-289.

[4] P. Cortez, J. Teixeira, A. Cerdeira, et al. Using data mining for wine quality assessment [C] International Conference on Discovery Science. Springer, Berlin, Heidelberg, 2009, pp.66-79.

[5] T. Larkin, D. McManus, An analytical toast to wine: Using stacked generalization to predict wine preference [J] Statistical Analysis and Data Mining: The ASA Data Science Journal, 2020, 13(5): 451-464.

[6] P. Schober, C. Boer, L. A. Schwarte, Correlation coefficients: appropriate use and interpretation[J]. Anesthesia & Analgesia, 2018, 126(5): 1763-1768.

[7] Y. Gupta, Selection of important features and predicting wine quality using machine learning techniques [J] Procedia Computer Science, 2018, 125: 305-312.

[8] D. Borkin, A. Némethová, G. Michaľčonok, et al. Impact of data normalization on classification model accuracy [J] Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 2019, 27(45): 79-84.

[9] J. S. Simonoff, Smoothing methods in statistics [M] Springer Science & Business Media, 2012.

[10] O. Sagi, L. Rokach, Ensemble learning: A survey [J] Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1249.

[11] W. S. Sarle, Neural networks and statistical models, In Proceedings of the Nineteenth Annual SAS Users Groups International Conference, Cary, NC, SAS Institute, Inc. 1994, pp. 1538–1550.

[12] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system [C] Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp.785-794.

[13] L. Breiman, Random forests [J] Machine learning, 2001, 45(1): 5-32.

[14] W. S. Noble, What is a support vector machine? [J] Nature biotechnology, 2006, 24(12): 1565-1567.