A new red wine prediction framework using machine learning

A new red wine prediction framework using machine learning

Chao Ye[1,a], Ke Li[2,b], Guo-zhu Jia[3,c*]

[1]College of physical and Electronics Engineering, Sichuan Normal University, Chengdu, China

[2]College of physical and Electronics Engineering, Sichuan Normal University, Chengdu, China

[3]College of physical and Electronics Engineering, Sichuan Normal University, Chengdu, China

[a] sa06@gxq-changzhi.com

[a]375704114@qq.com, [b]lk_1275@163.com, [c]jia1689500@126.com

Abstract—Red wine has become an integral part of people's lives and culture. Modeling the red wine quality is crucial. We propose a new framework to predict the red wine quality ratings. MF-DCCA was utilized to quantitatively investigate the cross-correlation between quality and physicochemical data. The long-range correlations importance was ranked. We compared two machine learning algorithms with other common algorithms implemented on the red wine data set, which was taken from UC Irvine Machine Learning Repository to ensure the reliability and performance. These data sets contain 1599 instances for red wine with 11 features of physicochemical data. Our model has better performance than other results.

1. INTRODUCTION

With the enlargement of red wine market, rapid optimization wine quality has become an integral part of red wine industry [1]. Currently manufacture promotes its products by using product certification, which requires a human expert to evaluate it, it is a time-consuming and expensive process [2]. Computer modeling and physical measurement are vital for the red wine quality, in order to improve the efficiency and accuracy of the prediction, some previous works devoted to wine ratings by machine algorithm(table I) and physical and chemical detection, such as direct injection mass spectrometry analysis combined with multivariate statistics [3], NIR Spectroscopy combined chemometrics[4],HNMR spectroscopy combined with multivariate statistical analysis[5], FIR; Genetic Fuzzy System [6].Furthermore, sensory tests were implemented, such as the relationship between the length of the tasting note and price [7], human sensory scores [8], age [9] as well as wine review [10].

However, a few researches are being performed on quantitatively analyzing the cross-correlation between physicochemical data and quality of red wine within multiple dimensions. In this paper, we first utilized the MF-DCCA model, which is a powerful tool for depicting the red wine quality due to that it's a robust method to quantitatively analyze cross-correlation in one or higher dimension, which is proposed by Zhou et al. [11]. Since then, MF-DCCA is used to explore the multifractal properties of cross-correlation between two time series [12,13]. Recently, the cross-correlations between individual investor sentiment and Chinese stock market return [14], online sentiment proxies [15], price and volume in European carbon futures markets [16] are checked. Due to the machine learning algorithms (SVM, NN, ID3, etc.) are time-consuming and the highest accuracies of them are less than 90%,

Boosting algorithms can be utilized to reduce bias and variance from the dataset, which have been widely used as empirical models for engineering applications. After the emergence of the gradient boosting machine [17], then Stochastic Gradient Boosted Decision Trees (GBDT) is proposed [18]. As an improvement of GBDT, XGBoost is widely applied in classification, regression and sorting problems due to its excellent predictive performance and sparse data processing capacity [19,20]. Moreover, LightGBM is also successfully applied in the fields of currency [21,22], transportation [23,24] and disease [25,26]. Hence, we propose a new framework combined MF-DCCA with LightGBM and XGBoost to classify red wine quality ratings.

2. MATERIALS AND METHODS

2.1 Data source and description

The data selected red wine data from the Wine Quality Data Set from the University of Minho, which has a total of 1599 samples, contains physicochemical data red and wine quality. The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). Each entry denotes a given test (analytical or sensory) and the final database was exported into a single sheet (.csv), the description of data attributes as table II. The red wine physicochemical data don't follow normal distribution (table III and Fig.1). The trend plots of 11 physicochemical data are as shown in Fig.1, which fluctuate violently. The red wine quality ratings range between 0 (very bad) and 10 (excellent).

2.2 Framework Model

The framework of predicting red wine quality ratings is shown in Fig.2, including the dataset, data preprocessing unit, data analysis unit, and data prediction unit and data prediction evaluation unit, which are detailed in the next sections.

2.3 Methods

2.3.1 MF-DCCA

In this paper, we adopt MF-DCCA to investigate the dynamic relationship between 11 physicochemical data of red wine and its quality. The red wine record is regarded as pseudo-time series in data set. Now we introduce the MF-DCCA method briefly.

For two time series, x (t) and y (t) (t = 1, 2, …, N), where N is the length of the two time series. The MF-DCCA method can be conducted as follow:

Firstly. Construct the profile of the time series

$$X(i) = \sum_{k=1}^{i} (x(k) - \bar{x}) \tag{1}$$

$$Y(i) = \sum_{k=1}^{i} (y(k) - \bar{y}) \tag{2}$$

Where $\bar{x}$ and $\bar{y}$ are the average of two time series respectively.

Secondly, the two profiles *X(i)* and *Y(i)* are divided into *Ns=int[N/s*] non-overlapping segments with the same length *s*. Because the length of the time series (N) is not always an integral multiple of the time scale s, a small fraction of each profile may remain. In order to avoid the loss of information contained in the time series, the same procedure is repeated starting from the end of each profile. In this way, 2*Ns* non-overlapping segments are obtained together. Here we set *8<s<N/4*.

Thirdly, for each segment λ of the profile, we calculate the local trends by an mth-order polynomial fit.

$$\hat{X}_{\lambda}(i) = \hat{a}_k i^m + \ldots \hat{a}_l i + \hat{a}_0 \tag{3}$$

$$\hat{Y}_{\lambda}(i) = \hat{b}_k i^m + \ldots \hat{b}_l i + \hat{b}_0 \tag{4}$$

Where $i = 1,2,\ldots,s$; $\lambda = 1, 2,\ldots 2N_s; m = 1,2,..$

Fourthly, we calculate a least-squares fit of each segment λ,and get the detrended covariance.

$$F^2(s,\lambda) = \frac{1}{s}\sum_{k=1}^{s}\{X[(\lambda-1)s+k]-\hat{X}_\lambda(k)\} \times\{Y[(\lambda-1)s+k]-\hat{Y}_\lambda(k)\} \quad (5)$$

For each segment $\lambda = 1,2,\ldots,N_s$ and

$$F^2(s,\lambda) = \frac{1}{s}\sum_{k=1}^{s}\{X[N-(\lambda-N_s)s+k]-\hat{X}_\lambda(k)\} \times\{Y[N-(\lambda-N_s)s+k]-\hat{Y}_\lambda(k)\} \quad (6)$$

For each segment $\lambda = N_s+1, N_s+2,\ldots, 2N_s$..

Fifthly, we obtain the *qth* order fluctuation function by squaring and averaging fluctuations over all segments.

$$F_q(s) = \{\frac{1}{2N_s}\sum_{\lambda=1}^{2N_s}[F^2(s,\lambda)]^{q/2}\}^{1/q} \quad (7)$$

For any real value $q \neq 0$.

When $q = 0$

$$F_0(s) = \exp\{\frac{1}{4N_s}\sum_{\lambda=1}^{2N_s}\ln[F^2(s,\lambda)]\} \quad (8)$$

Sixthly, by observing the log–log plots of $F_q(s)$ versus s for each value of q, we can analyze the scale behavior of the fluctuation function. If two series are long-range cross-correlated, we can obtain a power-law expression:

$$F_q(s) \sim s^{H_{xy}(q)} \quad (9)$$

It is equivalent to

$$\log F_q(s) = H_{xy}(q)\log(s) + \log C \quad (10)$$

Here the scaling exponent$H_{xy}(q)$, known as the Generalized Hurst Exponent, can be derived by observing the slope of the log–log plots of F_q (s) versus s through the method of ordinary least squares (OLS). In this paper, q varies from −10 to 10. If $q > 0$, $H_{xy}(q)$ reveals the scaling behaviors of the segments with large fluctuations and if $q < 0$, $H_{xy}(q)$ reveals the scaling behaviors of the segments with small fluctuations. Especially, when $q = 2$, MF-DCCA reduces to DCCA, the scaling exponent $H_{xy}(q)$ reduces to the Hurst exponent (H = $H_{xy}(q=2)$) which varies from 0 to 1. When $H_{xy}(q) > 0.5$, the cross-correlations between the two time series related to q are persistent (positive), suggesting that rises in one series are statistically likely to be followed by the increases of the other one. When $H_{xy}(q) < 0.5$, the cross-correlations between the two time series related to q are anti-persistent (negative), indicating that increases in one series are statistically likely to be followed by the decreases of the other one. When $H_{xy}(q) = 0.5$, the two time series are not cross-correlated with each other, x (t) has no effect on y (t), and vice versa.

Following Yuan [27], ΔH is calculated to measure the degree of multifractality.

$$\Delta H = H_{max}(q) - H_{min}(q) \quad (11)$$

Here, a large ΔH denotes a strong degree of multifractality, and vice versa.

Following Shadkhoo and Jafari [28], the Renyi exponent $\tau_{xy}(q)$ can be used to characterize the multifractal characteris-tic.

$$\tau_{xy}(q) = qH_{xy}(q) - 1 \quad (12)$$

If the Renyi exponent function is a nonlinear function of q, the cross-correlation of *x(t)* and *y(t)* is multifractal, otherwise, it is monofractal.

With Legendre transform, the cross-correlation can be described by the singularity strength α_{xy} *(q)* and the singularity spectrum f_{xy} *(α)*,

$$\alpha_{xy}(q) = \tau'_{xy}(q) = H_{xy}(q) + qH'_{xy}(q) \quad (13)$$

$$f_{xy}(q) = q\alpha_{xy}(q) - \tau_{xy}(q) = q(\alpha_{xy}(q) - H_{xy}(q)) + 1 \quad (14)$$

where $\tau'_{xy}(q)$ and $H'_{xy}(q)$ denote the derivative of $\tau_{xy}(q)$ and $H_{xy}(q)$ respectively. The singularity strength $\alpha_{xy}(q)$ characterizes singularities and monofractality in the time series. The singularity spectrum $f_{xy}(q)$ describes the singularity content of the time series. The width of the spectrum $\Delta\alpha_{xy}$, which is equivalent to $max(\alpha_{xy}) - min(\alpha_{xy})$, describes the strength of multifractality.

2.3.2 XGBoost algorithm

Compared with the traditional Gradient Boosting decision tree, XGBoost has a great improvement. According to [29], the objective function is as follows:

The target function is shown in the Eq. (15) is composed of error and regularization item. θ is the hyperparameter of the objective function. Where, l is the error between the predicted value and the target value. ΥT represents spanning tree pruning to prevent over-fitting. It can further prevent the model from over-fitting and improve its generalization ability.

$$F_{Obj}(\theta) = L(\theta) + \Omega(\theta) = l(\hat{y}_i, y_i) + \gamma T + \frac{1}{2}\lambda\|w\|^2 \quad (15)$$

Eq. (16) can be obtained from the addition model. Then, the tree structure is established to minimize the target. It learns from the output tree and the previous residuals. S_tT_i characterize the tree generated by the instance in the iteration. Eq. (16) becomes very complex in solving function problems. Thus, Eq. (17) increases the possibility of solving the problem through the taylor expansion. Finally, the optimal objective function is obtained by greedy algorithm [30].

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \bar{y}_i^{t-1} + S_t(T_i)) + \Omega(\theta) \quad (16)$$

$$L(\theta) = \sum_{i=1}^{n} [l(y_i, \bar{y}_i^{t-1}) + g_i S_t(T_i) + \frac{1}{2} h_i S_t^2(T_i)] + \Omega(\theta) \quad (17)$$

Where, $g_i = \partial_{\bar{y}_i^{t-1}} l(y_i, \bar{y}_i^{t-1})$ and $h_i = \partial^2_{\bar{y}_i^{t-1}} l(y_i, \bar{y}_i^{t-1})$. (18)

2.3.3 LightGBM algorithm

LightGBM is an effective algorithm for solving classification, sorting and regression task [26]. It takes up less memory and gets better prediction [31] than XGBoost. Its calculation flow is similar to XGBoost, and the advantages are as follows. Based on histogram selection decision tree algorithm [32], it speeds up the training process and reduces memory consumption. Moreover, Combine advanced network communication, the model can optimize parallel learning to improve speed. At the same time, it uses the leaf wise strategy to find the leaf with the maximum allocator gain, which is more effective than the XGBoost. It can reduce the error during training to increase the accuracy, but it also increases the risk of over-fitting, which can be limited by setting the maximum depth.

3. RESULTS AND DISCUSSION

3.1 Cross-correlation Test

In order to check the cross-correlation between physico-chemical indicators of red wine and its quality, we introduce a new cross-correlation test developed by Podobnik et al. [33], a qualitative method for testing the significance level of the cross- correlation between any two time series.

For two time series, $\{x_t, t=1,...,N\}$ and $\{y_t, t=1,\ldots,N\}$, where N is the equal length of these two series, the test statistic is

$$Q_{cc}(m) = N^2 \sum_{k=1}^{m} \frac{C_i^2}{N-i} \tag{19}$$

Where the cross-correlation function is

$$C_i = \frac{\sum_{k=i+1}^{N} x_k y_{k-i}}{\sqrt{\sum_{k=1}^{N} x_k^2 \sum_{k=1}^{N} y_k^2}} \tag{20}$$

The cross-correlation statistic *Qcc (m)* is approximately χ^2 (m) distributed with m degrees of freedom. For comparison, we set the critical values for the χ^2 (m) distribution at the 5% level of significance for different degrees of freedom m. If the cross-correlation statistic Q_{cc} *(m)* exceeds the critical values of χ^2 *(m)* ($Q_{cc}(m) > \chi^2_{0.95}$ (m)), the cross-correlations between the two series are significant; if $Q_{cc}(m)$ is below the critical value of χ^2 (m), this suggests that the two series are not cross-correlated.

In Fig.3, with the degrees of freedom varying from 10^0 to 10^3, we find that the cross-correlation statistics $Q_{cc}(m)$ between physicochemical data and quality are always larger than the critical values for the $\chi^2(m)$ distribution at the 5% level of significance. Thus, we can overall reject the null hypothesis of no cross-correlations. In other words, the long-range cross-correlations exist between the analyzed physicochemical data and quality of wine.

3.2 Multifractal detrended Cross-correlation Analysis

In order to further study the cross-correlation between 11 physicochemical data and mass, we adopted MF-DCCA from the perspective of multifractal.

Column (2-12) in Table IV show the scaling cross-correlation exponents H_{xy} for red wine's quality and its each physicochemical data. When $q=2$, H_{xy} displays more than 0.5 for all columns, which demonstrates that there exists persistent between 11 physicochemical data and quality. H_{xy} of volatile acidity (0.6695) is the largest of all the data. That is to say, volatile acidity has more influence on quality than other physicochemical data. For further, we compare the degree of multifractality (ΔH) in each column. The ΔH of residual sugar (1.0471) is the largest of all the data, indicating that the multifractality of the cross-correlation between quality and residual sugar is the largest of all the data.

We compared the cross-correlation index between 11 physicochemical data and mass throughout the sampling period, with q varying from -10 to 10. The results are shown in Table IV. From the longitudinal contrast of each column, whether q takes small values or large values, the generalized cross-correlation exponents are larger than 0.5, indicating that cross-correlations of small fluctuations and large fluctuations are persistent. In addition, if the exponent H_{xy} is constant, then the series is a single fractal; Otherwise, it's multifractal. The cross-correlation index of all scales decreases with the change of Q between -10 and 10, indicating that the cross-correlation between the 11 physicochemical data and mass is highly multiple. Fig.4 displays the image of the statistics in Table IV.

According to Eq.(12), the Renyi exponent $\tau_{xy}(q)$ is estimated(Fig.5). $\tau_{xy}(q)$ is nonlinearly dependent on q, which provides further evidence in support of multifractality. For further research, we use the multifractal spectrum width in Figure 6 to test the relationship between the two series. The multifractal spectrum of 11 physical and chemical data and quality is not displayed as a single point, indicating that there is a multifractal relationship between 11 physical and chemical data and quality. We also calculated the width of the multifractal spectrum $\Delta\alpha$. The results are presented in Table V. The $\Delta\alpha$ of

residual sugar is 1.2681, larger than that of quality and other data, suggesting that the multifractality degree of the cross-correlation between quality and residual sugar is stronger than that of the cross-correlation between quality and other data. The conclusion is consistent with the results of ΔH as above.

3.3 Rolling Windows Analysis

Rolling windows is capable to capture the dynamical cross-correlation over time. The length of each window is fixed at 256 records. The rolling step is 1 record. We calculate out the scaling exponent for the two pairs of series in each window for $q = 2$. The graphical representations are shown in Fig.7. As we can see from Fig.7, the dynamic scaling exponents are overall larger than 0.5, indicating the very strong persistent cross-correlations between 11 physicochemical data and quality.

3.4 Classification Algorithm Analysis

Compared with other algorithm implemented on the same data, our framework empirical results have a higher accuracy with accuracy of 91.04% (table VI and table VII). It can attribute to the optimization of machine learning algorithms. LightGMB made some optimization, using the histogram subtraction to make an acceleration, leaf-wine leaf growth strategy with depth limitation to reduce errors and get better accuracy, adding decision rules for category features to avoid computational and memory overhead by converting features to multidimensional one-hot feature [24]. XGBoost was following advantages, on a single platform 10 times faster than existing methods, using all the cores of machine ,allowing the use of wide variety of computing environments, helping in avoiding data over-fitting, being equipped to detect and deal with missing values ,being a flexible classifier because of it gives the user the option to set objective function as desired etc [34]. The confusion matrix as well as Roc Curves of XGBoost and LightGBM algorithm are shown in Fig 8-11, respectively.

4. CONCLUSION

Correct understanding of red wine physicochemical properties is the basis and premise for red wine quality success. We proposed a new framework combined MF-DCCA with XGBoost and LightGBM. For the correlation importance and classification results that is obtained, We think that the proposed approach is an advance in red wine quality classification. Residual sugar contributes the most complexly to red wine quality while the weakest cross-correlation are volatile acidity and chlorides, respectively. Both LightGBM and XGBoost achieved higher classification accuracy than the other machine learning algorithms.

The relative importance of the inputs brought significant views regarding the impact of the analytical tests. Since some physicochemical variables can be controlled in the production process this information can be used to improve the red wine quality. The proposed new framework is based on objective tests and thus it can be integrated into a decision support system, improving the speed and quality of the oenologist performance. Hence, we expect this will help to promote red wine in research and development processes.

5. FIGURES AND TABLES

5.1 Tables

TABLE I. THE MACHINE ALGORITHMS IN RED WINE RESEARCHES

Authors	***Methods***
Cortez, Cerdeira, Almeida, Matos, & Reis, 2009 [35]	MR,SVM,NN
Gupta, 2018 [2]	LR,NN,SVM
Appalasamy, Mustapha, Rizal, Johari, & Mansor, 2012[36]	ID3,Naïve Bayesian
Er & Atasoy, 2016 [37]	RF,SVM,k-NN
Beltre A¡n et al., 2006 [38]	NN,Bayesian
Ahammed & Abedin, 2018 [39]	LDA,MLR,RF
Our study	XGBoost,LightGBM

TABLE II. THE DESCRIPTION OF DATA ATTRIBUTES

Attributes	***Description***
Fixed Acidity (FA)	Most of the acids associated with wine are either fixed or non-volatile (not easy to evaporate)
Volatile Acidity (VA)	Too much acetic acid in wine can cause an unpleasant vinegar taste
Citric Acid (CA)	A small amount of citric acid can increase the freshness and flavor of wine
Residual Sugar (RS)	The residual sugar after fermentation has stopped is rarely found in wines below 1 gram per liter, and wines above 45 grams per liter are considered sweet
Chlorides(C)	The amount of salt in the wine
Free Sulfur Dioxide (FSD)	The free form of SO2 exists in the equilibrium state between molecular SO2 (as a dissolved gas) and bisulfite ion; it can prevent the growth of microorganisms and the oxidation of wine
Total Sulfur Dioxide (TSD)	The amount of free and bound forms of SO 2; At low concentrations, there is almost no SO2 detected in wine, but when the free SO2 concentration exceeds 50 ppm, SO2 will be evident in the aroma and taste of wine.
Density(D)	The density of water is close to the density of water, which depends on the alcohol and sugar content
PH(PH)	Describe the acidity or alkalinity of wine, from 0 (very acid) to 14 (very alkaline); most wines have a pH between 3-4
Sulphates(S)	A wine additive that can increase the content of sulfur dioxide gas (S02), it is an antibacterial and antioxidant
Alcohol(A)	Alcohol content of wine
Quality(Q)	Output variables (based on sensory data) (scores between 0 and 10

TABLE III. THE DESCRIPTIVE STATISTICS FOR PHYSICOCHEMICAL DATA AND QUALITY

Attributes	***Mean***	***Min***	***Max***	***Std***	***Skew***	***Kurt***	***P_value***
FA	8.3	4.6	15	1.7	0.9	1.13	8.1E-75
VA	0.5	0.1	1.5	0.2	0.6	1.2	3.06E-48
CA	0.2	0	1	0.1	0.3	-0.7	1.29E-15
RS	2.5	0.9	15.5	1.4	4.5	28.6	0
C	0.1	0.01	0.6	0.04	5.6	41.7	0
FSD	15	1	72	10.4	1.2	2.0	1.0E-149
RSD	46	6	289	32.8	1.5	3.8	0
D	0.9	0.9	1.0	0.01	0.07	0.9	1.83E-13
PH	3.3	2.7	4.0	0.15	0.1	0.8	3.61E-12
S	0.6	0.3	2	0.1	2.4	11.7	0
A	10	8.4	14.9	1.0	0.8	0.2	4.4E-44
Q	5.6	3	8	0.80	0.2	0.2	0.0001

TABLE IV. CROSS-CORRELATION EXPONENTS BETWEEN 11 PHYSICOCHEMICAL DATA AND QUALITY

α	***FA***	***VA***	***CA***	***RS***	***C***	***FSD***	***TSD***	***D***	***PH***	***S***	***A***
-10	1.3	1.3	1.3	1.5	1.3	1.3	1.3	1.4	1.3	1.3	1.4
-9	1.3	1.3	1.3	1.5	1.3	1.3	1.3	1.4	1.3	1.3	1.3
-8	1.3	1.3	1.3	1.5	1.3	1.3	1.3	1.3	1.3	1.3	1.3
-7	1.3	1.3	1.3	1.4	1.3	1.3	1.3	1.3	1.3	1.3	1.3

-6	1.3	1.3	1.3	1.4	1.3	1.3	1.2	1.3	1.2	1.3	1.3
-5	1.3	1.3	1.2	1.4	1.3	1.3	1.2	1.3	1.2	1.3	1.3
-4	1.2	1.2	1.2	1.4	1.2	1.2	1.2	1.3	1.2	1.3	1.2
-3	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.2	1.2	1.2	1.2
-2	1.2	1.2	1.2	1.2	1.1	1.1	1.1	1.2	1.1	1.2	1.1
-1	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0	0.6	0.6	0.6	0.7	0.6	0.6	0.6	0.6	0.6	0.6	0.6
1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
2	0.6	0.6	0.6	0.6	0.5	0.5	0.6	0.6	0.6	0.6	0.6
3	0.6	0.6	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.6	0.6
4	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.6	0.6
5	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.6	0.5
6	0.5	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.6	0.5
7	0.5	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.5
8	0.5	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.5
9	0.5	0.6	0.6	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5
10	0.5	0.6	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Δ H	0.8	0.7	0.7	1.0	0.8	0.8	0.7	0.8	0.7	0.7	0.8

Note: The meaning of each abbreviation sees table II.

TABLE V. 11 MULTI-FRACTAL DEGREE AND MULTI-FRACTAL SPECTRUM WIDTH OF CORRELATION BETWEEN PHYSICAL AND CHEMICAL DATA AND QUALITY

α	***FA***	***VA***	***CA***	***RS***	***C***	***FSD***
α_{max}	1.487	1.418	1.402	1.633	1.476	1.4426
α_{min}	0.498	0.585	0.540	0.365	0.438	0.461
$\Delta\alpha$	0.988	0.832	0.862	1.268	1.037	0.9816
α	TSD	D	PH	S	A	
α_{max}	1.405	1.504	1.412	1.434	1.513	
α_{min}	0.511	0.493	0.506	0.519	0.500	
$\Delta\alpha$	0.894	1.010	0.906	0.915	1.012	

Note: The meaning of each abbreviation sees table II.

TABLE VI. PERFORMANCE RESULTS OF THE CLASSIFICATION OF RED WINE SAMPLE QUALITIES

Test model classifier	***Recall（%）***	***Precision（%）***	***F1 measure（%）***	***Accuracy（%）***
LightGBM	85.63	86.67	86.15	91.04
XGBoost	90.67	90.67	90.67	91.04

TABLE VII. THE COMPARISON OF THE ACCURACY

Authors	***Model***		***Model accuracy(%)***
Cortez et al., 2009	Multiple regression		59.1
	support vector machine		59.1
	Neural networks		62.4
Er & Atasoy, 2016	cross validation	Support Vector Machines	0.54
		k-Nearest Neighbourhood	58.3
		Random Forests	64.8
	percentage split	Support Vector Machines	69.6
		k-Nearest Neighbourhood	59.1

		Random Forests	65.6
	cross validation(after using PCA)	Support Vector Machines	71.9
		k-Nearest Neighbourhood	58
		Random Forests	64.8
	percentage split(after using PCA)	Support Vector Machines	71.2
		k-Nearest Neighbourhood	56.9
		Random Forests	67.8
Appalasamy et al., 2012	ID3		60
	Naïve bayesian		58.8
(Gupta, 2018)	all features	neural networks	83.05
		support vector machine	75.66
	selected features	neural networks	85.97
		support vector machine	81.03
Our study	XGboost,		91.04
	LightGBM		91.04

5.2 FIGURES

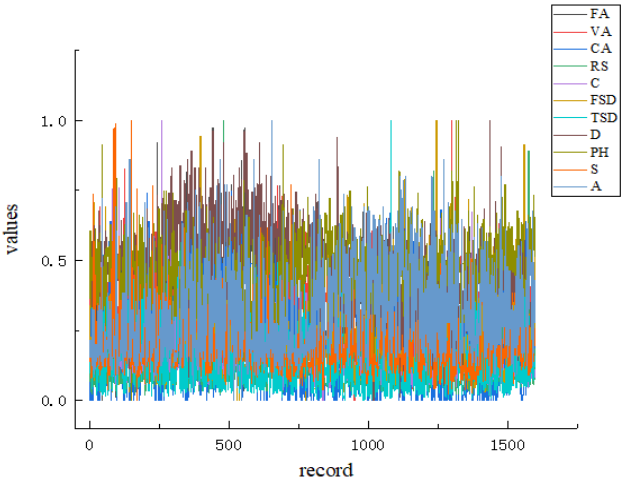

Figure 1. The trend plots of 11 physicochemical data.

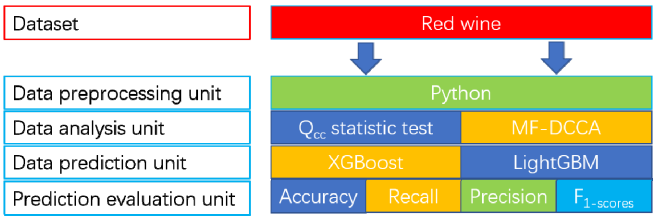

Figure 2. The framework of predicting red wine quality ratings by XGBoost and LightGBM algorithms.

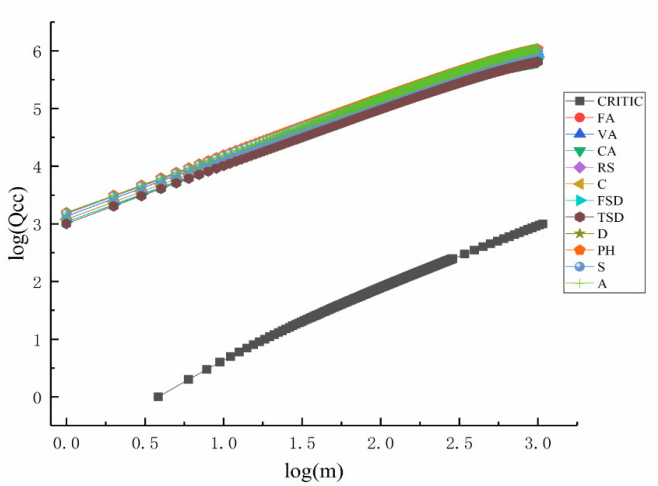

Figure 3. The cross-correlation statistics between 11 physicochemical data and quality.

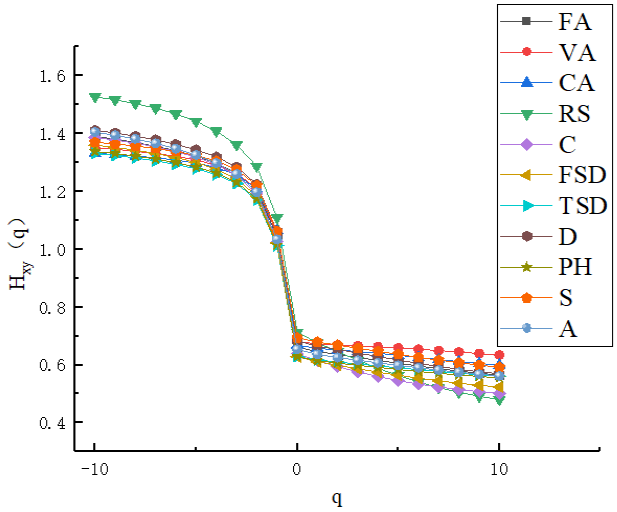

Figure 4. The cross-correlation exponents between 11 physicochemical data and quality

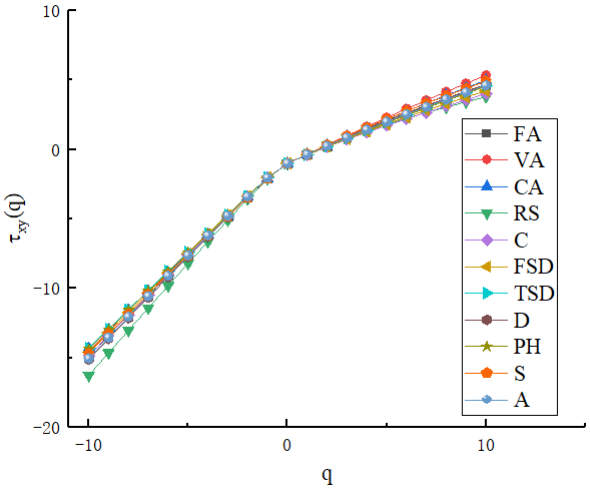

Figure 5. The relationship τxy(q) and q of 11 physicochemical data and quality.

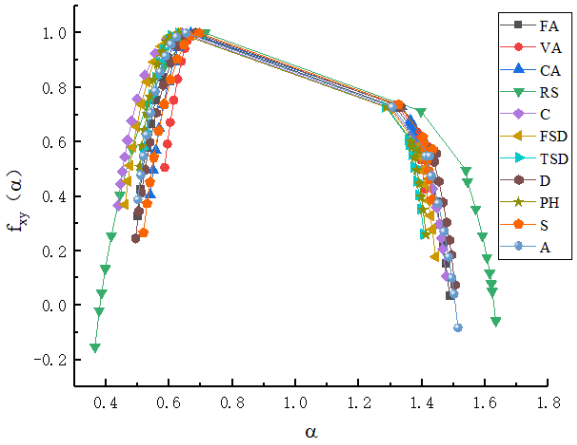

Figure 6. Multifractality degree and multifractal spectra widths for the correlation between 11 physicochemical data quality.

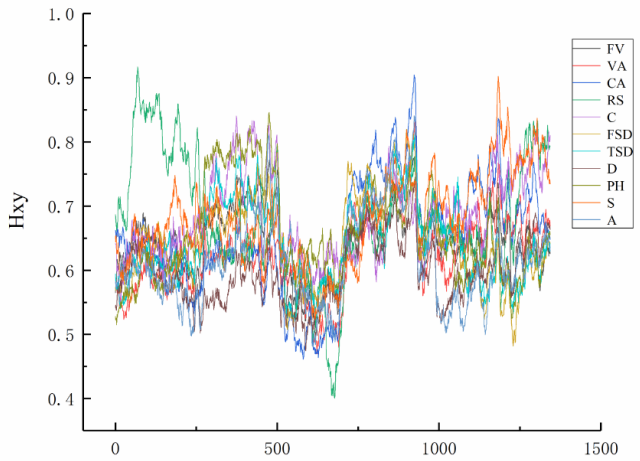

Figure 7. Scaling exponents for q = 2 for 11 physicochemical data and quality with window moving.

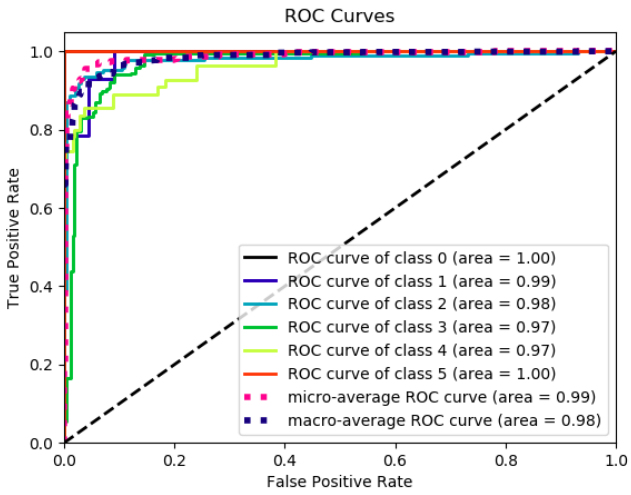

Figure 8. The Roc Curves of XGBoost algorithm.

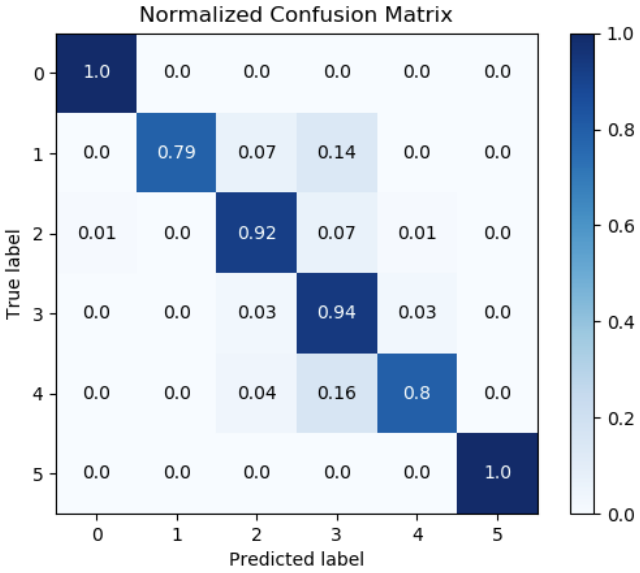

Figure 9. The confusion matrix of XGBoost algorithm.

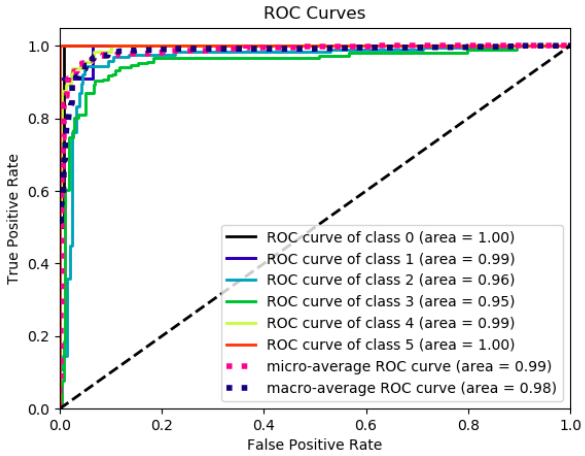

Figure 10. The Roc Curves of LightGBM algorithm.

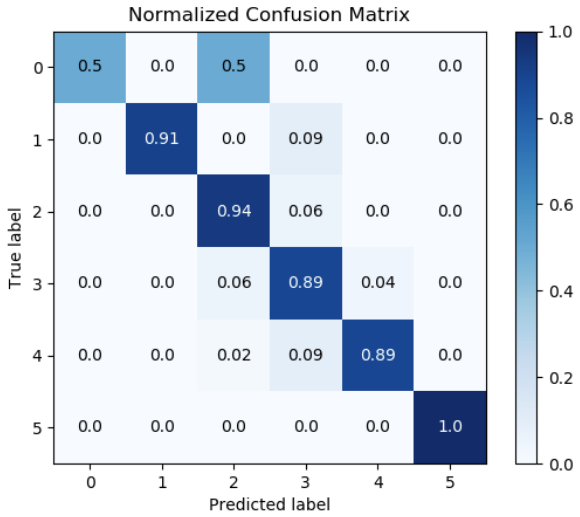

Figure 11. The confusion matrix of LightGBM algorithm.

REFERENCES

[1] Netzel, M., Strass, G., Bitsch, I., Ke A¶nitz, R., Christmann, M., & Bitsch, R. (2003). Effect of grape processing on selected antioxidant phenolics in red wine. Journal of Food Engineering, 56(2-3), 223-228.
[2] Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305-312.
[3] Villagra, E., Santos, L. S., Vaz, B. G., Eberlin, M. N., & Laurie, V. F. (2012). Varietal discrimination of Chilean wines by direct injection mass spectrometry analysis combined with multivariate statistics. Food chemistry, 131(2), 692-697.
[4] Shen, F., Yang, D., Ying, Y., Li, B., Zheng, Y., & Jiang, T. (2012). Discrimination between Shaoxing wines and other Chinese rice wines by near-infrared spectroscopy and chemometrics. Food and bioprocess technology, 5(2), 786-795.
[5] Fan, S., Zhong, Q., Fauhl-Hassek, C., Pfister, M. K.-H., Horn, B., & Huang, Z. (2018). Classification of Chinese wine varieties using 1H NMR spectroscopy combined with multivariate statistical analysis. Food Control, 88, 113-122.
[6] Nebot, e A. n., Mugica, F., & Escobet, A. (2015). Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques. Paper presented at the SIMULTECH.
[7] Ramirez, C. D. (2010). Do tasting notes add value? Evidence from Napa wines. Journal of Wine Economics, 5(1), 143-163.
[8] Legin, A., Rudnitskaya, A., Lvova, L., Vlasov, Y., Di Natale, C., & D.amico, A. (2003). Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. Analytica Chimica Acta, 484(1), 33-44.
[9] Yu, H., Lin, H., Xu, H., Ying, Y., Li, B., & Pan, X. (2008). Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy. Journal of agricultural and food chemistry, 56(2), 307-313.
[10] Chen, B., Le, H., Atkison, T., & Che, D. (2017). A Wineinformatics Study for White-box Classification Algorithms to Understand and Evaluate Wine Judges. Trans. MLDM, 10(1), 3-24.
[11] Zhou, W.-X. (2008). Multifractal detrended cross-correlation analysis for two nonstationary signals. Physical Review E, 77(6), 066211.
[12] He, L.-Y., & Chen, S.-P. (2011). Multifractal detrended cross-correlation analysis of agricultural futures markets. Chaos, Solitons & Fractals, 44(6), 355-361.
[13] Ma, F., Wei, Y., & Huang, D. (2013). Multifractal detrended cross-correlation analysis between the Chinese stock market and surrounding stock markets. Physica A: Statistical Mechanics and its Applications, 392(7), 1659-1670.
[14] Ruan, Q., Yang, H., Lv, D., & Zhang, S. (2018). Cross-correlations between individual investor sentiment and Chinese stock market return: New perspective based on MF-DCCA. Physica A: Statistical Mechanics and its Applications, 503, 243-256.
[15] Zhang, Z., Zhang, Y., Shen, D., & Zhang, W. (2018). The cross-correlations between online sentiment proxies: Evidence from Google Trends and Twitter. Physica A: Statistical Mechanics and its Applications, 508, 67-75.
[16] Zou, S., & Zhang, T. (2020). Multifractal detrended cross-correlation analysis of the relation between price and volume in European carbon futures markets. Physica A: Statistical Mechanics and its Applications, 537, 122310.
[17] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
[18] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.
[19] Chen, R.-C., Gallagher, L., Blanco, R., & Culpepper, J. S. (2017a). Efficient Cost-Aware Cascade Ranking in Multi-Stage Retrieval.

[20] Chen, R.-C., Gallagher, L., Blanco, R., & Culpepper, J. S. (2017b). Efficient cost-aware cascade ranking in multi-stage retrieval. Paper presented at the Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.
[21] Chowdhury, R., Rahman, M. A., Rahman, M. S., & Mahdy, M. (2019). Predicting and Forecasting the Price of Constituents and Index of Cryptocurrency Using Machine Learning. arXiv preprint arXiv:1905.08444.
[22] Sun, X., Liu, M., & Sima, Z. (2018). A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Research Letters.
[23] Mei, Z., Xiang, F., & Zhen-hui, L. (2018). Short-Term Traffic Flow Prediction Based on Combination Model of XgboostLightgbm. Paper presented at the 2018 International Conference on Sensor Networks and Signal Processing (SNSP).
[24] Wang, B., Wang, Y., Qin, K., & Xia, Q. (2018). Detecting Transportation Modes Based on LightGBM Classifier from GPS Trajectory Data. Paper presented at the 2018 26th International Conference on Geoinformatics.
[25] Chen, X., & Liu, X. (2018). A Weighted Bagging LightGBM Model for Potential lncRNA-Disease Association Identification. Paper presented at the International Conference on Bio-Inspired Computing: Theories and Applications.
[26] Wang, D., Zhang, Y., & Zhao, Y. (2017). LightGBM: an effective miRNA classification method in breast cancer patients. Paper presented at the Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics.
[27] Yuan, Y., Zhuang, X.-t., & Jin, X. (2009). Measuring multifractality of stock price fluctuation using multifractal detrended fluctuation analysis. Physica A: Statistical Mechanics and its Applications, 388(11), 2189-2197.
[28] Shadkhoo, S., & Jafari, G. (2009). Multifractal detrended cross-correlation analysis of temporal and spatial seismic data. The European Physical Journal B, 72(4), 679.
[29] Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018). XGBoost classifier for DDoS attack detection and analysis in SDN-Based cloud. Paper presented at the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp).
[30] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.
[31] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39.
[32] Jin, R., & Agrawal, G. (2003). Communication and memory efficient parallel decision tree construction. Paper presented at the Proceedings of the 2003 SIAM International Conference on Data Mining.
[33] Podobnik, B., Grosse, I., Horvati ¨ A‡, D., Ilic, S., Ivanov, P. C., & Stanley, H. E. (2009). Quantifying cross-correlations using local and global detrending approaches. The European Physical Journal B, 71(2), 243.
[34] Dhaliwal, S., Nahid, A.-A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. Information, 9(7), 149.
[35] Cortez, P., Cerdeira, A. n., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.
[36] Appalasamy, P., Mustapha, A., Rizal, N., Johari, F., & Mansor, A. (2012). Classification-based Data Mining Approach for Quality Control in Wine Production. Journal of Applied Sciences, 12(6), 598-601.
[37] Er, Y., & Atasoy, A. (2016). The classification of white wine and red wine according to their physicochemical qualities. International Journal of Intelligent Systems and Applications in Engineering, 23-26.

[38] Beltre A¡n, N., Duarte-Mermoud, M., Bustos, M., Salah, S., Loyola, E., Pee A?a-Neira, A., & Jalocha, J. (2006). Feature extraction and classification of Chilean wines. Journal of Food Engineering, 75(1), 1-10.
[39] Ahammed, B., & Abedin, M. (2018). Predicting wine types with different classification techniques. Model Assisted Statistics and Applications, 13(1), 85-93.