Wine Quality Prediction Based on Machine Learning Techniques

Yogesh Gupta and Amit Saraswat

Abstract Nowadays, it is extremely difficult to choose wines as there are numerous wine manufacturers. In response to the increase in customer base of wine, wine companies need to improve their quality and sales. There have been many attempts to develop a methodological approach for assessment of wine quality. In this paper, machine learning methods such as decision tree, random forest and support vector are used to check the quality of two types of wine: red and white. This work takes into account various ingredients of wine to predict its quality. The experiments show the superiority of random forest over decision tree and support vector classifiers.

Keywords Red wine · White wine · Decision tree · Random forest · Support vector

1 Introduction

Nowadays, all the industries are improving the quality of their products due to the huge competition in market. They use various technologies to enhance the production and to smooth the whole process. But, it becomes more expensive due to the increase in demands of product. This work explores the role of machine learning techniques to check the quality of wine. It is extremely difficult to choose wines as there are numerous wine manufacturers. In response to the increase in customer base of wine, wine companies need to improve their quality and sales. We have developed a method that considers various ingredients of wine to predict its quality using decision tree, random forest and support vector classifiers. We predict wine quality on a scale of 0–10. The main reason behind this is that the traditional methods involved in wine quality prediction are extremely time consuming and do not give a good accuracy;

Y. Gupta
Department of Computer Science, BML Munjal University, Gurugram, India
e-mail: yogesh.gupta@bmu.edu.in

A. Saraswat (✉)
Department of Electrical Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
e-mail: amit.saraswat@jaipur.manipal.edu

S. Dwivedi et al. (eds.), *Flexible Electronics for Electric Vehicles*, Lecture Notes in Electrical Engineering 863, https://doi.org/10.1007/978-981-19-0588-9_61

hence, there have been several attempts at reducing time and improving the accuracy. However, a trained and experienced wine quality tester (human) would provide better results but that is time consuming and costly.

In recent years, ingesting of wine has increased basically because it makes a positive impact on our health as it helps in keeping our heart rate in check. Consumption of wine in small amounts can prevent us from strokes. Increased consumption of wine has resulted in companies going for quality assessment tests. The main objective of this work is to forecast the wine quality taking some features as inputs. Higher the score out of 10, better is the quality. There are 12 input variables in total and one output, i.e. quality of wine.

The paper is planned as follows: Sect. 2 presents the existing work related to wine quality prediction. Section 3 discusses the methodology in detail. Experimental results and analysis are done in Sect. 4. At the last, conclusion and future directions are drawn in Sect. 5.

2 Related Work

There have been few works present in literature based on machine learning to predict wine quality. The main reason is that the traditional methods involved in wine quality prediction are extremely time consuming and do not give a good accuracy; hence, there have been several attempts at reducing time and improving the accuracy. After back propagation algorithm came into existence in 1974 [1], neural networks became very popular. Recently, support vector machine is introduced [2, 3], and it is gaining more attention nowadays. *SVM* and neural network are used in predictions due to their flexibility and learning capabilities. But, model selection and variable selection are the critical issues in applying these machine learning techniques. The selection of suitable variable [4] plays an important role not only in discarding irrelevant inputs, but it leads to a simpler model also. This simpler model can be easily interpreted and gives better performance. Although a simpler model has limited learning capabilities, complex models may over-fit the data.

Cortez et al. [5] assessed the wine quality using a regression approach which preserved the order of grades. Sensitivity analysis was given as knowledge to measure changes in response for variations in input variable. Moreover, Cortez et al. considered all features to predict wine quality. Cortez et al. [6] proposed an approach for predicting wine taste preferences of human being. They used three machine learning techniques: support vector machine, neural network and regression. Yin et al. [7] used soft measurement based on multivariate methods to evaluate wine quality. They included principal component regression, partial least squares regression, ordinary least squares regression and modified partial least squares regression in their approach. Ribeiro et al. [8] presented a model to predict organoleptic parameters from chemical parameters of the vinification process using machine learning techniques.

Gupta [9] developed neural network and support vector machine and evaluated the performances on wine data sets. Malik [10] proposed a prediction method based on the reviews of the customers. They created a database based on that information collected from reviewers. Further, they identified the parameters to perform classification based on that information. Pawar et al. [11] predicted the wine quality using regression and support vector classifiers. Reddy and Govindarajulu [12] presented a new prediction model for wine quality based on a centric clustering approach. They used red wine data set for performing classification. Further, they assigned different weights to all attributes using Gaussian distribution process. The obtained results were satisfactory.

3 Methodology

In this work, white and red wine data sets [13] are used containing a total of 6497 records. The data includes values of 12 input variables and one output variable quality (on a scale of 1–10).

Figure 1 shows the processes adopted in methodology. First the collected data is pre-processed. Data pre-processing includes two different stages: data cleaning and data transformation. In data cleaning process, missing data and noisy data are handled, whereas normalization is done in data transformation process. Subsequently, some of the important features are selected. Further, the quality of wine is predicted, and accuracy is computed for all developed models.

In this work, decision tree, random forest and support vector classifiers are used. These classifiers are supervised learning algorithm and use labelled data. Decision tree is a tool which is used in determining what result will be obtained if a certain decision is made. Each decision will lead to another set of options and the option we choose will determine what result we will obtain. An example where the logic of decision trees can be seen in coding would be if–else, switch statements. In a decision tree, classification is based on different procedures such as recursive binary splitting and then identifying the cost of a split.

Random forest classifier is an extension of the decision tree. Every single decision tree would be producing a class prediction. The idea which governs the working of

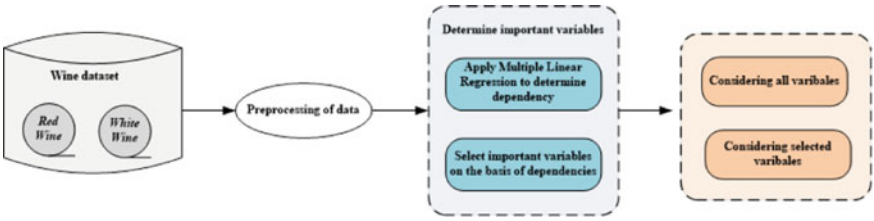

Fig. 1 Block diagram of used methodology

a random forest classifier is that if a large number of uncorrelated models start operating as a committee, then it would always outperform any of the individual constituent models. Support vector classifier is used to classify different classes of data points in any data set. The objective of support vector is to find a hyperplane on an N-dimensional space. The hyperplane should have a maximum margin, so we can easily classify additional data which we might add in our data at some point of time in future with convenience.

4 Result Analysis and Discussion

The performances of used algorithms are evaluated in terms of accuracy at certain error tolerances (ACC_t) and mean absolute deviation (*MAD*). First, data is divided into test and training data. In this work, 20% of the data is used as testing data. After that, normalization is performed on data as algorithms may perform poorly if individual inputs, i.e. features do not look like normally distributed. Normalization fits the values in a specific range. Further, decision tree, random forest and support vector classifier are applied and obtained accuracy and MAD values as given in Tables 1 and 2 for both types of wine.

Table 1 summarizes the values of *MAD* and ACC_t for *decision tree, random forest* and *support vector* in case of red wine data set. It is clear from this table that *random forest* outperforms other algorithms. Similarly, Table 2 presents the results of both the metrics *MAD and* ACC_t for white wine data set. This table clearly shows that

Table 1 Performance of red wine quality prediction models. Estimation metrics include (ACC_t) and (*MAD*)

	Decision tree	Random forest	Support vector machine
MAD	0.537	0.513	0.583
$ACC_t = 0.25$	0.286	0.302	0.206
$ACC_t = 0.50$	0.363	0.481	0.317
$ACC_t = 1.0$	0.525	0.677	0.489
$ACC_t = 2.0$	0.682	0.769	0.615

Table 2 Performance of white wine quality prediction models. Estimation metrics include (ACC_t) and (*MAD*)

	Decision tree	Random forest	Support vector machine
MAD	0.583	0.571	0.618
$ACC_t = 0.25$	0.237	0.275	0.196
$ACC_t = 0.50$	0.428	0.491	0.386
$ACC_t = 1.0$	0.549	0.657	0.512
$ACC_t = 2.0$	0.688	0.778	0.635

random forest outperforms decision tree and support vector with higher accuracy at low error tolerance values and less mean absolute deviation.

5 Conclusions and Future Work

In the last few years, the usage of machine learning has been increased in wine industry for quality prediction. In this direction, wine quality certification plays a very important role. This paper explores the usage of three different types of machine learning techniques to predict the quality. The experiments show the superiority of random forest over decision tree and support vector classifiers. The reason of this superiority of random forest is the differences in training phases. The training in random forest always guarantees an optimum fit the data. In future, large data set can be taken for experiments, and other machine learning techniques may be explored.

References

1. Werbos P (1974) Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University, Cambridge, MA
2. Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, pp 144–152
3. Smola A, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222
4. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(7):1157–1182
5. Cortez P, Teixeira J, Cerdeira A, Almeida F, Matos T, Reis J (2009) Using data mining for wine quality assessment. In: Proceeding of international conference of discovery science. Lecture notes on artificial intelligence, pp 66–79
6. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009) Modeling wine preferences by data mining from physicochemical properties. Decis Support Syst 47:547–553
7. Yin S, Liu L, Gao X, Karimi H (2014) Multivariate methods based soft measurement for wine quality evaluation. Abstr Appl Anal 2014:1–7
8. Ribeiro J, Neves J, Sanchez J, Delgado M, Machado J, Novais P (2009) Wine vinification prediction using data mining tools. Comput Comput Intell 78–85
9. Gupta Y (2018) Selection of important features and predicting wine quality using machine learning techniques. Proc Comput Sci 125:305–312
10. Malik MSI (2020) Predicting users' review helpfulness: the role of significant review and reviewer characteristics. Soft Comput 24:13913–13928
11. Pawar D, Mahajan A, Boithe S (2019) Wine quality prediction using machine learning. Int J Comput Appl Technol Res 8(9):385–388
12. Reddy YS, Govindarajulu P (2017) An efficient user centric clustering approach for product recommendation based on majority voting: a case study on wine data set. Int J Comput Sci Netw Secur 17(10):103–111
13. https://archive.ics.uci.edu/ml/machine-learning-databases/wine/