Wine Quality Analysis Using Machine Learning

Bipul Shaw, Ankur Kumar Suman and Biswarup Chakraborty

Abstract Almost from the beginning of mankind, there has been the existence of different kinds of wine. It has also become very important for us to know the quality of the wine, before consuming it. In the last few decades, the food industry has grown enormously and so are the food quality analysis and its "rating" process. We often come across cases in which a consumer falls sick because of consuming low-quality food, so it has become a necessary evil for us to have a quality analysis of a product before selling the product, "evil" because it adds up extra cost to the production of the final product. Similarly, it is also necessary to do a quality analysis of wine and there have been different methods used to determine the quality of the wine, but we often get confused regarding which method to rely on! This paper focuses on the comparative study over different classification algorithms for wine quality analysis which are: SVM, random forest and multilayer perceptron and to know which of the above-mentioned classification algorithms give more accurate result.

Keywords Machine learning · Support vector machine · Random forest · Multilayer perceptron · Wine quality

1 Introduction

As we all know, that heavy research is being done in the field of machine learning, but surprisingly it is not a new topic! It was started back in 1950 by Alan Turing. He proposed a "learning machine" that could learn and become artificially intelligent,

B. Shaw (✉) · A. K. Suman · B. Chakraborty
Department of Computer Science and Engineering, Institute of Engineering
and Management, Saltlake, India
e-mail: bipulshaw2014@gmail.com

A. K. Suman
e-mail: aksankuraks@gmail.com

B. Chakraborty
e-mail: biswarup.98.kkr@gmail.com

J. K. Mandal and D. Bhattacharya (eds.), *Emerging Technology in Modelling and Graphics*, Advances in Intelligent Systems and Computing 937,
https://doi.org/10.1007/978-981-13-7403-6_23

which marked the starting of the era of machine learning! But even though it was proposed so early, it has gained an enormous interest growth in last few years. The main four reasons for this sudden change are: rise in computing power, heavy amount of data generation, growth of deep learning and the rise of digital era. We can broadly classify machine learning into three subtopics: supervised learning, unsupervised learning and reinforcement learning. In supervised learning, we train the algorithm with the training data sets which includes both inputs and labels (targets or outputs). After training the algorithm with a lot of data sets, we try to create the logic for predicting the labels for the new data. There are two types of supervised learning: classification and regression. Then comes unsupervised learning in which training data sets does not include targets; here, we do not tell the system where to go, but the system has to understand itself from the data we give because here training data is not structured. It also has two types of method, namely clustering and anomaly detection. Reinforcement learning differs in the aspect of input/output pairs which may need not be presented and suboptimal actions need not be explicitly corrected. Instead of that, the focus is given upon performance, which involves balance between exploration and exploitation.

Our food industry has also become a large-scale user of these techniques, and in that we have a small section of wine quality analysis in which machine learning is used extensively. The reason for the sudden growth in wine quality check is because it has a direct relation with our health, and moreover it helps us to check the variation in the condition of heart. Another valid reason for this can be the increase in the amount of wine consumption and forcing their respective companies to have assessment on wine quality and grading certification to maintain their name and survive in the corporate world. And hence, the above-stated reasons motivated us to write a paper on wine quality analysis and to compare the result of different algorithms.

Similarly, we can also state that since the quality of a wine does not depend on a single factor but on multiple attributes so it becomes comparatively easier to check the quality of a wine by machine learning rather than the human tasters! Moreover, it also helps us to know that which physical/chemical attribute is affecting the quality of the wine in which way (either positive or negative). In this paper, we have implemented and compared the results of three techniques, namely support vector machine, random forest and multilayer perceptron to check the quality of wine.

2 Literature Survey

There have been many researches done on topics like “wine quality analysis”, “price prediction of wines” and “conditions in favour of preparing wine”. Some of the earlier works on these topics are as follows: in [1], the authors associated wine drinking with increase in heart rate variability in women along with coronary heart disease. In [2], the authors tell the wine applications with the help of electronic noses. In [3], the authors used Gaussian regression process and multitask learning to predict the price of wine. They used past data of wine price to predict the price of wine in future. They concluded that advanced machine learning technique has the potential to predict the price of wine. In [4], the authors mentioned that the price and quality of wine depend on the weather in which the grapes were cultivated. He derived a price equation using several factors. In [5], the authors predicted wine verification using data mining tools. In [6], the authors analysed the quality of wine using a decision tree and other tools. In [7], the authors model the wine preferences using data mining from some physiochemical properties. In [8], the authors introduce us to the elimination of recursive features with random features for PTR-MS analysis of agro-industrial products. In [9], the authors extract rules from multilayer perceptron in some classification problems. It is a clustering-based approach. In [10], the authors compare multivariate analytic techniques. They also listed out the pros and cons of recursive partitioning analysis. In [11], the author shows us an efficient algorithm to generate classification rules. In [12], the author has done the wine quality analysis using several classification approaches with different feature sets like principal component analysis, recursive feature elimination and nonlinear decision tree. In [13], the author has also classified wine quality with imbalanced data using synthetic minority over-sampling technique (SMOTE), decision tree, adaptive boosting and random forest. And hence after being encouraged by the previous works as stated above, we applied more than one ML algorithms to predict the quality of the wine, namely SVM, random forest and multilayer perceptron.

3 Methodologies

See Fig. 1.

3.1 Data Preparation

The data set contains 4898 instances of red wine from the UCI machine learning repository. The physical properties which are in the data set are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol and finally quality. There are physical as well as chemical variables that influence the quality of wines. Tartaric acid, citric acid and malic acid are present in the wine, whereas ascorbic and sulphurous acids are added during the winemaking. Residual sugar determines the sweetness of the wine. Although it is not the only factor which determines the sweetness of the wine, it still plays the major role in determining the taste of the wine. And due to yeast metabolism, there is generation of alcohol by which wine gets its alcoholic properties.

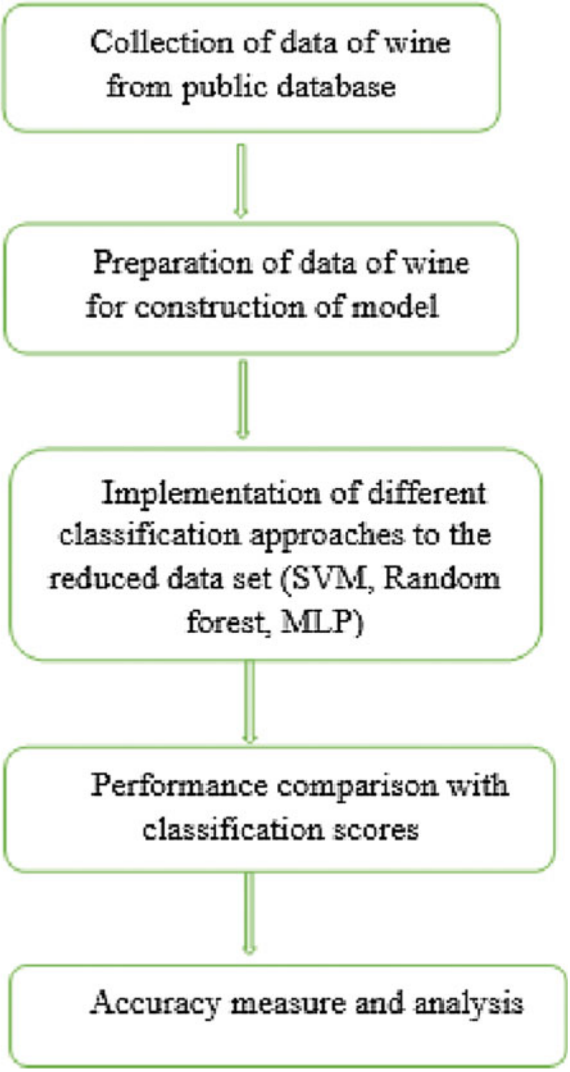
Collection of data of wine from public database
Preparation of data of wine for construction of model
Implementation of different classification approaches to the reduced data set (SVM, Random forest, MLP)
Performance comparison with classification scores
Accuracy measure and analysis

Fig. 1 Flow chart showcasing our working structure in this paper

3.2 Selection of Algorithms

In this paper, our motive was to analyse the quality of wine using supervised learning method and to be more specific, by different classification techniques. So in order to remain on our motive, we decided to take three algorithms in supervised learning method, but each one of its kind. So, we took SVM as first algorithm, random forest as second algorithm and multilayer perceptron as our third algorithm. So basically, we will be comparing among a simple classifier such as SVM, an algorithm which works on the principle of decision tree such as random forest and a classifier which uses artificial neural network like multilayer perceptron to see which type of classifier gives best result in our stated problem.

3.3 Comparative Study

In this paper, we have used three different kinds of techniques, namely: SVM, random forest and multilayer perceptron, to find and predict the quality of the given wine data with the help of created logic and by training the data sets. First of all, we will implement the SVM algorithm to the test data set and calculate its result, after this we will implement the random forest algorithm to the test data set and calculate its result, and finally we will implement the multilayer perceptron algorithm and calculate its result as well. After calculating the results, we will make a bar graph for better distinction between the performances of the three algorithms used in this paper.

Support Vector Machine

Support vector machine or SVM is a supervised learning model with associated learning algorithms that analyses data used for classification and regression analysis. Its basic approach is to separate the positive and negative class with the largest margin. It is based on Vapnik statistical learning theory [14]. It has many good properties like margin maximization, high fitting accuracy, small number of tuneable parameters and kernel technology adopted in high-dimensional feature space. We can also say that it is a non-probabilistic binary linear classifier.

Random Forest

Random forest is a supervised learning method which consists of random decision tree which works in coordination with classification and regression. They operate by constructing multiple layers of decision tree at the time of training and output the classes or mean prediction of the individual trees. Now by taking the probability of all the decision trees into account, the overall probability is calculated. It is basically used for the data sets having high dimensionality where the individual variables are non-stationary and highly noisy.

Multilayer Perceptron

Multilayer perceptron is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and nonlinear activation distinguish MLP from a linear perceptron. It can even distinguish data that are not linearly separable. For now, MLP classifier supports only cross-entropy loss function, which allows probability estimates by running predict_proba method.

3.4 Performance Measure Metrics

The parameters used to determine the quality of the wine by the above-mentioned algorithms are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates and alcohol content. In this paper, the parameter used to compare the performance and validations of the classifiers is accuracy score, whereas according to the definition, accuracy in machine learning is the number of correct predictions made, divided by the total number of prediction.

4 Result

There are several solutions for the wine quality analysis as well as several ways to do it. The above-stated problem has been solved by various algorithms but which one gives the best result? We have considered three basic algorithms which were used to determine the quality of the wine and studied that which algorithm gives the best possible result. In this paper after dividing the data set into two groups, namely training data set and test data set, we trained each classifier based on the training data set and tested their (classifier's) efficiency on the test data set. So, each classifier is able to show the performance metrics, i.e. accuracy based on the test data set. We made a bar graph plot for better understanding of the comparative study of the classifiers based on the accuracy parameter. And hence, we were able to see that random forest algorithm gave the result with best accuracy and SVM with the worst among the three. The classification scores of SVM are given in Table 1, classification scores of random forest are given in Table 2, and classification scores of multilayer perceptron are given in Table 3. And a bar graph is shown in Fig. 2.

The following given graph is based on the comparative study among the three algorithms we used in this paper, namely SVM, random forest and multilayer perceptron by parameterizing their accuracies.

Table 1 Classification score for SVM

Quality	Precision	Recall	F1-score	Support
3	0.00	0.00	0.00	9
4	0.00	0.00	0.00	76
5	0.60	0.65	0.63	669
6	0.56	0.75	0.64	960
7	0.57	0.21	0.31	349
8	0.00	0.00	0.00	82
Average/total	0.53	0.57	0.53	2145

Accuracy score: 0.5729603729603729

Table 2 Classification score for random forest

Quality	Precision	Recall	F1-score	Support
4	0.98	0.46	0.62	140
5	0.84	0.86	0.85	1469
6	0.77	0.90	0.83	1846
7	0.91	0.67	0.77	730
8	1.00	0.39	0.56	111
9	1.00	0.20	0.33	5
Average/total	0.83	0.82	0.81	4352

Accuracy score: 0.8196231617647058

Table 3 Classification score for multilayer perceptron

Quality	Precision	Recall	F1-score	Support
4	0.79	0.56	0.65	140
5	0.85	0.79	0.82	1469
6	0.75	0.87	0.80	1846
7	0.78	0.68	0.72	730
8	0.88	0.40	0.55	111
9	1.00	0.20	0.33	5
Average/total	0.83	0.82	0.81	4352

Accuracy score: 0.787812548529

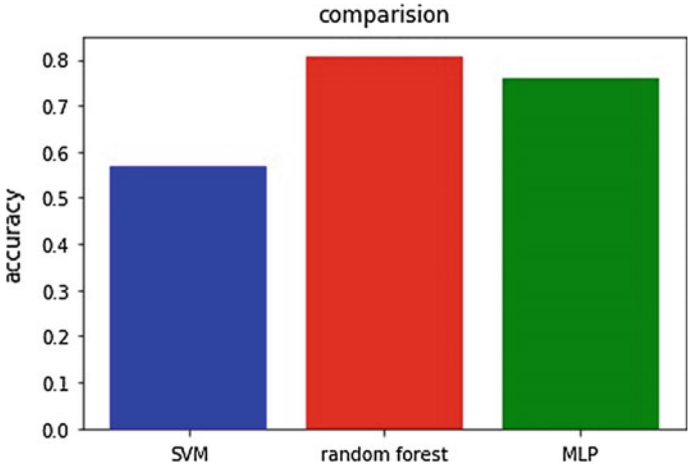

Fig. 2 Bar graph showing the comparison among the used algorithms, on the basis of accuracy

5 Conclusion

In this paper, our basic motive was to see that which of the three algorithms, SVM, random forest and MLP, give the best accurate result in the analysis of wine quality. During our comparative study between those algorithms, we came to know that random forest algorithm gives the best result with an accuracy percentage of 81.96, then comes multilayer perceptron algorithm with an accuracy percentage 78.78 and lastly comes the support vector machine algorithm with an accuracy percentage of 57.29. The reason why random forest algorithm gives the best result is because when a node is split during the formation of tree, the split that is chosen is no longer the best spilt among all the features. Instead, the split chosen is the best and due to this randomness, the prejudice of the system increases and hence yields an overall better model. Finally, we can conclude our paper with a result that random forest algorithm is the best among the above-stated algorithms and a question that, “Is there any other classification algorithm which can give better accuracy result than random forest algorithm?”.

References

1. I. Janszky, M. Ericson, M. Blom, A. Georgiades, J.O. Magnusson, H. Alinagizadeh, S. Ahnve, Wine drinking is associated with increased heart rate variability in women with coronary heart disease. Heart **91**(3), 314–318 (2005)
2. V. Preedy, M.L.R. Mendez, Wine applications with electronic noses, in *Electronic Noses and Tongues in Food Science* (Academic Press, Cambridge, MA, USA, 2016) pp. 137–151
3. M. Yeo, T. Fletcher, J. Shawe-Taylor, Machine learning in fine wine price prediction. J. Wine Econ. **10**(2), 151–172 (2015)
4. O. Ashenfelter, Predicting the quality and prices of Bordeaux wine. J. Wine Econ. **5**(1), 40–52 (2010)
5. J. Ribeiro, J. Neves, J. Sanchez, M. Delgado, J. Machado, P. Novais, Wine vinification prediction using data mining tools, in *Conference Proceedings, Computing and Computational Intelligence* (Tbilisi, Republic of Georgia, 2009, June) pp. 78–85
6. S. Lee, J. Park, K. Kang, Assessing wine quality using a decision tree, in *IEEE International Symposium on Systems and Engineering (ISSE)* (2015, Sept) pp. 176–178
7. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst., Elsevier **47**(4), 547–553 (2009)
8. P.M. Granitto, C.L. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr. Intell. Lab. Syst. **83**(2), 83–90 (2006). https://doi.org/10.1016/j.chemolab.2006.01.007
9. E.R. Hruschka, N.F. Ebecken, Extracting rules from multilayer perceptrons in classification problems: a clustering-based approach. Neurocomputing **70**(2), 384–397 (2006)
10. E.F. Cook, L. Goldman, Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J. Chronic Dis. **37**(9–10), 721–731 (1984)
11. S. Vijayarani, M. Divya, An efficient algorithm for generating classification rules. Int. J. Comput. Sci. Technol. **2**(4), 512–515 (2011)
12. S. Aich, A.A. Al-Absi, K.L. Hui, J.T. Lee, M. Sain, A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques, in *International Conference on Advanced Communication Technology* (2018), pp. 139–143

13. G. Hu, T. Xi, F. Mohammed, H. Miao, Classification of wine quality with imbalanced data, in *IEEE International Conferrence on Industrial Technology (ICIT)* (2016), pp. 1712–1717
14. V.N. Vapnik, The Nature of Statistical Learning Theory. New York: Berlin: Springer (1995)