Semestral work, Course CSC522: Automated Learning and Data Analysis, North Carolina State University, Raleigh, 2014

Prediction of wine quality from physiochemical properties

Bhashwanth Kadapagunta, Mahalaxmi Sanathkumar, Nitish Pandey, Parul Upadhyaya,Venkatesh Sambandamoorthy

***Abstract*-The purpose of this project is to model the quality of wines based on the physiochemical properties of the wine. The physiochemical test measures the presence of various chemicals such as alcohol, sulfur dioxide, chlorides, citric acids etc. Certain physical aspects of wines are also measured such as density in the tests. A large dataset containing almost 6500 instances of these physiochemical tests was used in order to model the quality of the wines. Various data mining techniques were used in order to classify the wines. Techniques such as neural networks, random forests and support vector machines showed the most promising results. Techniques such as k-nearest neighbors were also applied but the results were not as good as those of the other techniques. The models generated by these methods can be used to predict the quality of wines.**

I. INTRODUCTION

The premise for this study is largely based on the efforts of the wine industry to quantify the quality of wines. In order to do so there are two types of tests conducted on wines: physiochemical tests and sensory tests. Our study deals with the physiochemical tests conducted on the wines. These tests and the associated results are used to model the quality of wine. The task of predicting the quality of wine has further implications in preventing the adulteration of wines, determining the prices of wines and their market value. The goal of the project is to use the available data and come up with a more accurate model which can predict the quality of wine better than the existing approaches.

In the subsequent sections, we have discussed our work to model the quality of wines. We have used multiple data mining techniques in order to predict the quality of wines for our dataset. The dataset that we have used has been taken from the UCI Machine learning repository [1].

The dataset has two subsets one for white wines and the other for red wines. The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Each sub dataset contains 12 attributes. All the values are numeric and continuous. The first 11 attributes are the physiochemical aspects of the wine such as: sulfates, pH, total sulfur dioxide, alcohol, volatile acidity, free sulfur dioxide, fixed acidity, residual sugar, chlorides, density, and citric acid. The last attribute in the dataset is the quality attribute. This acts as the class labels for the records in the dataset. Some attributes have a profound impact on the quality of the dataset as compared to others.

All the experiments in the project were performed on weka (Waikato Environment for Knowledge Analysis) which is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The classification techniques that we have used were K-nearest neighbors, neural networks, random forests and support vector machines. The wine quality of the training records is in the range of 0 to 10. The output of these classification techniques is in the form of accuracy of the determined classes as compared to the training set. In order to improve the accuracy of these methods we did some preprocessing. We altered various parameters of the above mentioned algorithms in order to generate a better model for the data that had higher classification accuracy.

II. BACKGROUND

Most of the existing works use regression techniques in order to build the model. We came across papers that have made use of these techniques in order to generate the model and find the main components of the dataset that contribute to the overall accuracy of the model. These papers made use of data mining techniques such as regression trees, linear regression etc. One such paper [3] discussed Wine quality prediction by applying regression tree algorithm using multiple parameters to the normalized data set. Another paper uses classic regression techniques like Neural Networks and Support Vector Machines due to their flexibility and non-linear learning capabilities [2]. They use variable selection to preprocess the data and discard irrelevant inputs. Their contribution included using a simultaneous variable and model selection scheme using sensitivity analysis. An accuracy of 86.8% was achieved using this approach. But this method of using sensitivity analysis introduces a dependency on the preset tolerance value that accepts responses which are within one of the two nearest class labels and hence cannot be considered as an accurate model. The existing work used only individual classifiers, but we improved the performance of the model by using ensemble classifiers. We also reduced the misclassification for the classes 5 and 6 which was not handled in previous works.

In recent times, the wine industry has put in a lot of investments in wine production and quality assessment. Among researchers, this has become an interesting topic to research on the wine evaluation techniques and to obtain best possible method or combination of methods that has high accuracy in determining the quality of wine. A lot of data mining techniques like classification and anomaly detection algorithms can be applied to the wine samples to classify wine based on several characteristics or to extract information about the best and worst wine. The paper that we referred discusses a similar concept [2]. They have modeled the wine type preferences on a continuous scale ranging from 0 to 10 (signifying tastes ranging from very bad to excellent taste). They have applied Support vector machine algorithm (SVM), multiple regression (MR) and Neural Networks (NN). They have given the consensus about their choice which is SVM because it has the best performance among all three methods. But they have also mentioned about the advantages that other methods have over each other. For example: Multiple regression is easier to interpret as compared to the other two methods. They have also mentioned about how the change in parameters or use of different functions in each of the algorithm affects the results. For example: if different hidden node or minimization cost functions are used, one can obtain different neural network results. Our study is based on a similar background. As mentioned in introduction section, we are making use of four algorithms (Neural Networks, Support vector machine, K-nearest neighbors and Random forests) with feature selection and subsequently model selection along with some pre-processing to arrive at an algorithm that provides best accuracy and classifies the wine samples best. We have also laid out the comparison between performances of each of the algorithms used. Our study, application of algorithms and explanation of experiments are mentioned in the succeeding sections.

III. OUR CONTRIBUTIONS

A. Data pre-processing and variable selection

The dataset contains all numeric and continuous values. The characteristics of both the datasets have been shown in Table 1.

	Data Characteristics	
	Red Wine	***White Wine***
Mean	8.32	6.855
Standard Deviation	1.741	0.844
Minimum	4.6	3.8
Maximum	15.9	14.2

Table 1 – Data Characteristics

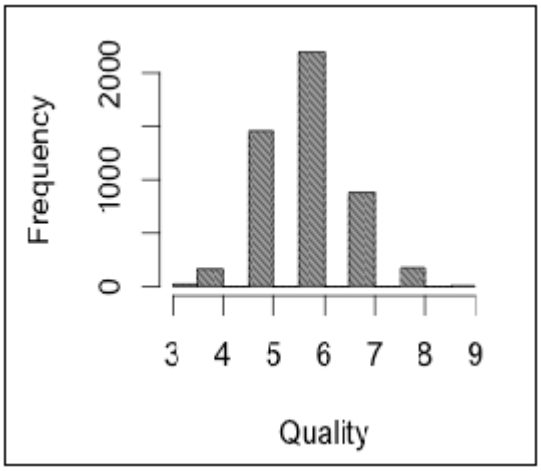


Fig 1 – white wine quality distribution

Fig 1 and Fig 2 were obtained by plotting the wine quality attribute in white wine and red wine respectively. We observed that the white wine dataset had no instances for quality labels 0, 1, 2 and 10. For the red wine dataset there were no instances for 0,1,2,9 and 10. Furthermore most of the data instances were classified for labels 5, 6 and 7 (Fig 2).

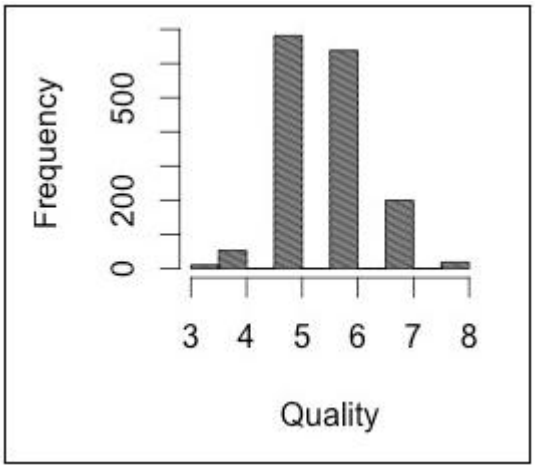
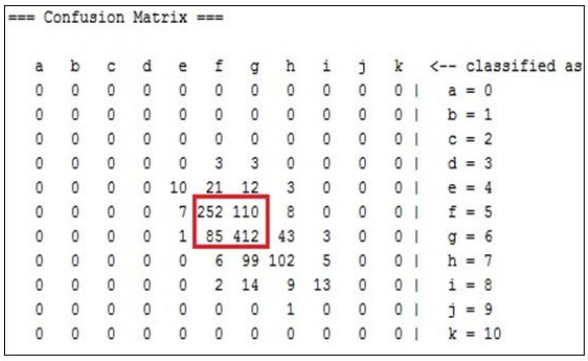


Fig 2 – red wine quality distribution

In order to reduce the variance in our dataset we standardized the dataset, we standardized all the attributes except the quality attribute in both of the datasets.



```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   <-- classified as
   0   0   0   0   0   0   0   0   0   0   0 |   a = 0
   0   0   0   0   0   0   0   0   0   0   0 |   b = 1
   0   0   0   0   0   0   0   0   0   0   0 |   c = 2
   0   0   0   0   0   3   3   0   0   0   0 |   d = 3
   0   0   0   0  10  21  12   3   0   0   0 |   e = 4
   0   0   0   0   7 252 110   8   0   0   0 |   f = 5
   0   0   0   0   1  85 412  43   3   0   0 |   g = 6
   0   0   0   0   0   6  99 102   5   0   0 |   h = 7
   0   0   0   0   0   2  14   9  13   0   0 |   i = 8
   0   0   0   0   0   0   0   1   0   0   0 |   j = 9
   0   0   0   0   0   0   0   0   0   0   0 |   k = 10
```

Fig 3 – confusion matrix for raw data

Based on our preliminary tests on the data we found that there were multiple misclassifications for labels 5 and 6 and there was considerable overlap between the two labels, as shown in Fig 3. This was affecting the accuracy of our classifiers, in order to rectify this we merged the class labels 5 and 6.

B. Classification techniques

The first method we used was the neural networks, which are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from other observations after executing a process of so-called learning from existing data. We used the multilayer perceptron, varied the number of hidden layers and multiple output nodes according to the class labels. The learning rate was kept at 0.3 in order to prevent over fitting. The accuracy rate and the error rate were measured for all the variations of hidden layers.

The next method we used was Support Vector Machines, a technique motivated by statistical learning theory and is suitable for numerous classification tasks. We used the LibSVM package, varied the Gamma value, the Polynomial Degree and the Kernel Type parameters to determine the best fit for the model.

We then used Random Forest which is an ensemble learning method for classification that operates by constructing several decision trees at training time. The output class is the mode of the classes output by individual trees, thereby ensuring a higher accuracy than individual trees. The algorithm was tested by varying the number of trees, the number of features and the depth of trees to test the algorithm.

The last method that we used was k-nearest neighbors; a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. We used Euclidean distance as the similarity measure to find the nearest neighbors. We tested the algorithm for multiple values of K and measured the accuracy.

IV. EXPERIMENTS AND RESULTS

Two datasets containing the physiochemical properties of red and white wines were chosen for the project. For our experiments we divided the data into train and test. 75% of the data was used as training data and 25% of the data was used as the test data. All the techniques use the same distribution of data. This process was done for both the red wine dataset and white wine dataset. The results were measured in terms of the accuracy of these classifiers (Fig 5). The evaluation measures are tabulated as depicted in Table 5. Accuracy here is the number of correct classifications divided by the total number of classifications. We then compared the results to that of the original dataset. We used 10 fold cross validation to evaluate our results.

A. Results of Neural Networks

We ran the test for both original data and modified data and varied the no of hidden layers between 3, 4 and 10. The best value of accuracy was achieved for hidden layers at three.

B. Results of Support vector machines

In case of white wine, the best accuracy was obtained by using the Radial Kernel Type and a Gamma value of 2.2, whereas for red wine, the best accuracy was obtained by using the Radial Kernel Type and a Gamma value of 0.8

C. Results of Random Forest

The model built on standardized data yielded best results for both white and red wines when the number of features was 1, the depth of the tree was 0 (unlimited) and the number of trees was 20. The accuracies are as shown in Fig4.

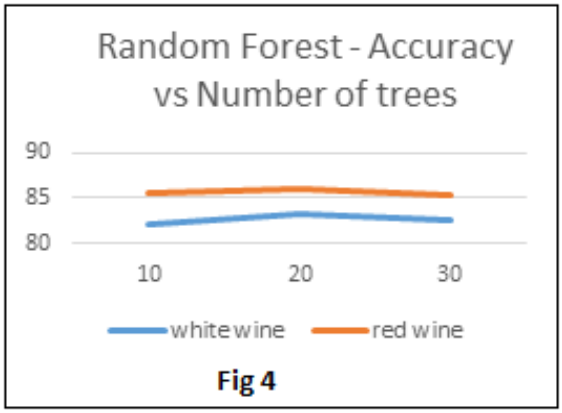


Fig 4

	KNN		Neural Network		SVM		Random Forest	
	White	Red	White	Red	White	Red	White	Red
Accuracy	75.4902	76.7157	77.124	82.8643	82.107	85.5	83.1669	86
RMSE	0.1859	0.181	0.178	0.15	0.18	0.159	0.1566	0.1425

Table 2 – Evaluation measures of techniques used

D. Results of K-nearest neighbors

The experiment conducted on both the datasets using K-nearest neighbors algorithm yielded the best accuracy for a value of k=5

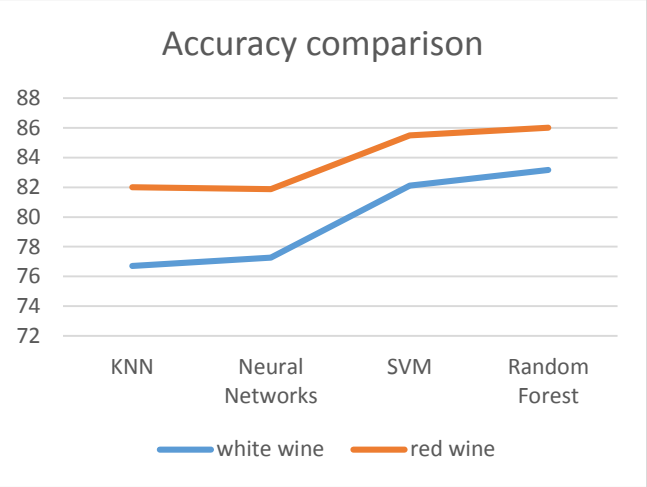


Fig 5 – Accuracy comparison

E. Validation

To validate the algorithms so as to gauge how effectively it performs on our data set, we used 10 fold Cross Validation and results from the confusion matrix. The validation results using 10 fold cross validation are tabulated in Table 3.

	White Wine	Red Wine
KNN	77.011	82.8018
Neural Networks	76.3373	82.8643
Support Vector Machines	82.5	85.25
Random Forest	84.1568	86.429

Table 3 – cross validation results

Observing the confusion matrix, we found that the original dataset had a lot of cross classification in the classes 5 and 6 due to the overlapping of characteristics. When classes 5 and 6 were merged, the misclassifications reduced drastically and the accuracy using Random Forest went up from 67.6048% to 86%

V. CONCLUSIONS

Wine quality prediction is of much interest among wine industries to improve production process of wine and its sales [2]. Data mining techniques can be used to predict such wine qualities by extracting relevant attributes from the raw data. After preprocessing the data and merging of two classes with similar characteristics, we applied techniques like Random Forest and Support Vector Machine on the dataset. Based on our tests and the results from [2] we observe that the Random Forest method provided the best classification for both red wine and white wine datasets with accuracies of 86% and 83.1699% respectively. The merging of classes with labels 5 and 6 considerably increased the accuracy of our classifier as compared to the accuracy of the raw data. The number of misclassifications between class5 and class 6 was brought down to 0 as depicted in Fig 6.

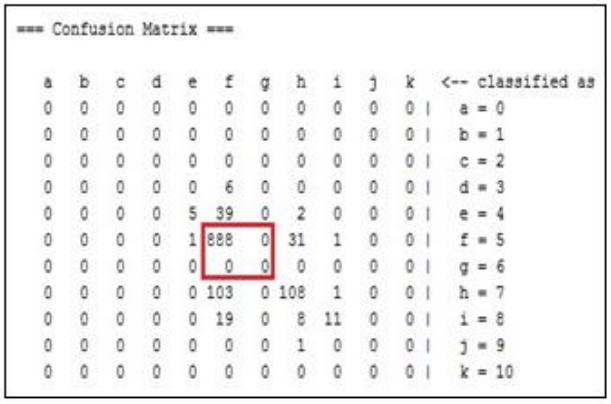


Fig 6 – confusion matrix of final data

The superior performance of the Random Forest method, an ensemble technique, can be attributed to the fact that it outputs the class that is the mode of the classes output by individual trees in the forest.

VI. FUTURE SCOPE

The prediction of the quality of wine can be further enhanced by improving the accuracy of the classifier. This can be attained by ensemble learning using a combination of classifiers. Boosting techniques can also be used to improve the performance of the classifier.

REFERENCES

[1] UCI Machine learning repository [http://archive.ics.uci.edu/ml/datasets/Wine+Quality]

[2] P.Cortez, A.Cerdeira, F.Almeida, T.Matos and J.Reis, "Modeling wine preferences by data mining from physicochemical properties," In Decision Support Systems 47 (2009) 547–533

[3] M.Horak," Prediction of wine quality from physiochemical properties," Semestral work, Course 336VD: Data Mining, Czech Technical University in Prague, 2009/10