International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

The Classification of White Wine and Red Wine According to Their Physicochemical Qualities

Yesim Er*[1], Ayten Atasoy[1]

Accepted 3rd September 2016

Abstract: The main purpose of this study is to predict wine quality based on physicochemical data. In this study, two large separate data sets which were taken from UC Irvine Machine Learning Repository were used. These data sets contain 1599 instances for red wine and 4898 instances for white wine with 11 features of physicochemical data such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH. First, the instances were successfully classified as red wine and white wine with the accuracy of 99.5229% by using Random Forests Algorithm. Then, the following three different data mining algorithms were used to classify the quality of both red wine and white wine: k-nearest-neighbourhood, random forests and support vector machines. There are 6 quality classes of red wine and 7 quality classes of white wine. The most successful classification was obtained by using Random Forests Algorithm. In this study, it is also observed that the use of principal component analysis in the feature selection increases the success rate of classification in Random Forests Algorithm.

Keywords: *Classification, Random Forests, Support Vector Machines, k-Nearest Neighbourhood*

1. Introduction

Today, varied consumers enjoy wine more and more. Wine industry is researching new technologies for both wine making and selling processes in order to back up this growth [1].

Physicochemical and sensory tests are used for evaluating wine certification [2]. The discrimination of wines is not an easy process owing to the complexity and heterogeneity of its headspace. The classification of wines is very important because of different reasons. These reasons are economic value of wine products, to protect and assure the quality of wines, to forbid adulteration of wines, and to control beverage processing [3].

Data mining technologies have been applied to classification of wine quality. The aim of machine learning methods similar to other applications is to create models from data to predict wine quality.

In the year of 1991, a "Wine" data set which contains 178 instances with measurements of 13 different chemical constituents such as alcohol, magnesium was donated into UCI repository to classify three cultivars from Italy [4]. For new data mining classifiers this data set has been majorly used as a benchmark because it is very easy to discriminate. For wine classification according to geographical region; principal component analysis (PCA) was carried out and reported [5]. The data they used in their study includes 33 Greek wines with physicochemical variables. Another work of wine classification depended on the physicochemical information. This information involved in wine aroma chromatograms as measured with a Fast GC Analyser [6]. In the latter study, three classification methods such as Linear Discriminant Analysis, Radial Basis Function Neural Networks, and Support Vector Machines (SVM) are compared according to their performance in a two-staged architecture.

Some have proposed a few applications of data mining techniques to wine quality assessment. Cortez *et al*. [1] proposed a taste prediction technique. Their taste prediction technique, a support vector machine, multiple regression, and a neural network were applied to chemical analysis of wines. Shanmuganathan's technique was about prediction the effects of season and climate on wine yields and wine quality [7]. The Wineinformatics system according to Chen *et al.* [8] idealized the flavour and characteristics of wine from natural-language reviews. They used hierarchical clustering and association rules.

2. Materials and Methods

2.1. Wine Data

The data set is a wine quality dataset that is publicly available for research purposes from https://archive.ics.uci.edu/ml/datasets/Wine+Quality [9]. Both dataset contains 1599 instances with 11 features for red wine and 4898 instances and the same 11 features for white wine. The inputs include objective tests (e.g. pH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine.

The features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic). Most wines are between 3-4 on the pH scale. Chloride is the amount of salt in the wine. Alcohol is the percent alcohol content of the wine.

The goal of the data set is to predict the rating that an expert will give to a wine sample, using a range of physicochemical

[1] *Karadeniz Technical University, Department of Electrical and Electronics Engineering, Trabzon-61080, Turkey*
** Corresponding Author: Email: yesim.er@ktu.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

properties, such as acidity and alcohol composition. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e. g. There is no data about grape types, wine brand, wine selling price, etc.).
Table 1 and Table 2 present the 11 different physicochemical properties and data statistics (minimum, maximum, mean, and standard deviation values of all instances for each feature) of white wine and red wine sets respectively.

Table 1. The physicochemical data statistics of white wine

Attribute (units)	Min	Max	Mean	StDv
Fixed acidity (g(tartaric acid)/dm^3)	3.800	14.20	6.855	0.844
Volatile acidity (g(acetic acid)/dm^3)	0.080	1.100	0.278	0.101
Citric acid (g/dm^3)	0.000	1.660	0.334	0.121
Residual sugar (g/dm^3)	0.600	65.80	6.391	5.072
Chlorides (g(sodium chloride)/dm^3)	0.009	0.346	0.046	0.022
Free sulfur dioxide (mg/dm^3)	2.000	289.0	35.31	17.01
Total sulfur dioxide (mg/dm^3)	9.000	440.0	138.4	42.50
Density (g/cm^3)	0.987	1.039	0.994	0.003
pH	2.720	3.820	3.188	0.151
Sulphates (g(potassium sulphate)/dm^3)	0.220	1.080	0.490	0.114
Alcohol (%vol)	8.000	14.20	10.51	1.231

Table 2. The physicochemical data statistics of red wine

Attribute (units)	Min	Max	Mean	StDv
Fixed acidity (g(tartaric acid)/dm^3)	4.600	15.90	8.320	1.741
Volatile acidity (g(acetic acid)/dm^3)	0.120	1.580	0.528	0.179
Citric acid (g/dm^3)	0.000	1.000	0.271	0.195
Residual sugar (g/dm^3)	0.900	15.50	2.539	1.410
Chlorides (g(sodium chloride)/dm^3)	0.012	0.611	0.087	0.047
Free sulfur dioxide (mg/dm^3)	1.000	72.00	15.87	10.46
Total sulfur dioxide (mg/dm^3)	6.000	289.0	46.47	32.89
Density (g/cm^3)	0.990	1.004	0.997	0.002
pH	2.740	4.010	3.311	0.154
Sulphates(g(potassium sulphate)/dm^3)	0.330	2.000	0.658	0.170
Alcohol (%vol)	8.400	14.90	10.42	1.066

2.2. Data Mining Approach

To evaluate performance of selected tool using the given dataset, several experiments are conducted. For evaluation purpose, two test modes are used, the k-fold cross-validation mode (k-fold cv) mode, and percentage split (holdout method) mode.
The k-fold cv refers to a widely used experimental testing procedure where the database is randomly divided into k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process repeated k times. At the end, the average value of the recorded measurement is found [10]. It is common to choose k as 10.
In percentage split mode, the database is randomly divided into two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge may be tested against the second set which is called testing set [10].
In this study, for k-fold cross-validation mode, different k values are tested for each method. The best classification results of each method are obtained when k value is chosen as 10.
Firstly, both datasets are separated into training and testing set by using 10-fold cross-validation method. Afterwards, both datasets are randomly divided into two groups called training and testing set. First set involves randomly 80% of dataset, and the other set involves the resting data.

2.3. Data Mining Techniques

In the original form of this datasets, two datasets were created, using red and white wine samples. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine.
First, these two datasets have been combined into one dataset to classify wine samples as red wine and white wine. Three different data mining algorithms were used in our study. Those classification algorithms applied on the data set are k-nearest neighbourhood (k-NN), random forests (RF), and support vector machines.

1) k-Nearest-Neighbourhood Classifiers: This method was depicted in the beginning of 1950s. Nearest-neighbourhood classifiers are depended on learning by analogy, this means a comparison between a test tuple with similar training tuples. The training tuples are described by n attributes. Each tuple corresponds a point in an n-dimensional space. All the training tuples are stocked in an n-dimensional pattern space. For an unknown tuple, a k-nearest-neighbourhood classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. k training tuples are called as the k "nearest neighbours" of the unknown tuple [11].

"Closeness" is a metric distance, likewise Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, ..., x_{1n})$ and $X_2 = (x_{21}, x_{22}, ..., x_{2n})$, is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (1)$$

2) Random Forests: This methodology uses a combination of tree predictors; each individual tree depends on a random vector. This random vector has identical and alike distribution for all trees in the forest. It was described by Breiman in 2001 [12].

3) Support Vector Machines: This method was derived from statistical learning theory by Vapnik and Chervonenkis, It was first introduced in 1992 by Boser, Guyon, and Vapnik. This method is used for the classification of both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. It searches for the linear optimal separating hyperplane in this new dimension. A hyperplane can separate data from two classes, with a suitable nonlinear mapping to sufficiently high dimension. The SVM uses support vectors and margins to find this hyperplane [11].

Then, the following three different data mining algorithms were used to classify the quality of both red and white wine samples: k-nearest-neighbourhood, random forests, and support vector machines.
Afterwards, principal component analysis in the feature selection was applied to the original both red wine dataset and white wine dataset for each method. Three data mining algorithms were used to classify the quality of both red wine samples and white wine samples.

3. Experimental Results

The three classification algorithms were used in our study to

classify the wine samples as red wine and white wine. A model was built using each method and applied to the wine data set. The classification results of the three classification algorithms are evaluated both test modes which are 10-fold cross-validation, and 80% percentage split.

Also, some of the standard performance measures (statistics) are calculated to evaluate the performance of the algorithms. The standard performance measures are recall, precision, F measure, and ROC values.

Table 3 presents the correctly classified instances results of the classification of red and white wine samples.

Table 3. Performance results of the classification of red and white wine samples

Test Modes	Classifier	Precision (%)	Recall (%)	F Measure (%)	ROC (%)
Cross Validation	SVM	99.1	99.1	99.1	98.6
	k-NN	99.2	99.2	99.2	99.0
	RF	99.5	99.5	99.5	99.8
Percentage Split	SVM	99.2	99.2	99.2	98.6
	k-NN	99.2	99.2	99.2	98.7
	RF	99.5	99.5	99.5	99.9

The most successful classification result of red and white wine samples was obtained by Random Forests Algorithm for both test modes. The accuracy of each cross-validation and percentage split mode with this algorithm is 99.5229%, and 99.4611% respectively.

Table 3 clearly shows that Random Forests algorithm outperforms from the other algorithms in two test modes.

For the classification of the quality of both red and white wine samples, the classification experiment is measured by the accuracy percentage of classifying the instances correctly into its class according to quality features which are ranged between 0 (very bad) and 10 (very excellent) as 11 different classes and totally 22 classes.

Table 4 and Table 5 present the correctly classified instances results of the classification of white and red wine sample qualities respectively.

For k-nearest-neighbourhood classifiers, different k values are tested for each test mode. When k value is increased, the achievement of the classification decreases. For this reason, k value is taken as 1.

Table 4. Performance results of the classification of white wine sample qualities

Test Modes	Classifier	Precision (%)	Recall (%)	F Measure (%)	ROC (%)
Cross Validation	SVM	39.6	52.1	44.1	66.7
	k-NN	65.1	65.4	65.2	75.0
	RF	71.0	70.4	69.5	87.3
Percentage Split	SVM	39.4	51.2	43.7	65.8
	k-NN	63.0	63.3	63.0	73.1
	RF	69.8	68.7	67.4	85.7

Table 5. Performance results of the classification of red wine sample qualities

Test Modes	Classifier	Precision (%)	Recall (%)	F Measure (%)	ROC (%)
Cross Validation	SVM	48.1	58.3	52.7	70.6
	k-NN	64.3	64.8	64.5	72.7
	RF	66.8	69.6	67.8	86.4
Percentage Split	SVM	49.7	59.1	53.9	70.8
	k-NN	65.6	65.6	65.5	72.8
	RF	69.6	71.9	70.5	87.2

The most successful classification result of white wine sample qualities as 11 classes was obtained by using Random Forest algorithm for both test modes. The accuracy of each cross-validation and percentage split mode with this algorithm is 70.3757%, and 68.6735% respectively.

The most successful classification result of red wine sample qualities as 11 classes was obtained by using Random Forest algorithm for both test modes. The accuracy of each cross-validation and percentage split mode with this algorithm is 69.606%, and 71.875% respectively.

Then, for increasing of classification success in this study, the number of features was reduced by using PCA algorithm and the process wine quality classification was repeated by using SVM, k-NN, and RF algorithms.

Table 6 and Table 7 present the correctly classified instances results of the classification of white and red wine sample qualities after applying PCA respectively.

Table 6. Performance results of the classification of white wine sample qualities after applying PCA

Test Modes	Classifier	Precision (%)	Recall (%)	F Measure (%)	ROC (%)
Cross Validation	SVM	39.8	52.2	44.0	66.3
	k-NN	64.5	64.7	64.6	74.4
	RF	70.7	69.9	68.8	86.9
Percentage Split	SVM	39.4	51.2	43.6	65.9
	k-NN	63.5	63.6	63.5	73.7
	RF	68.1	67.4	66.3	85.4

Table 7. Performance results of the classification of red wine sample qualities after applying PCA

Test Modes	Classifier	Precision (%)	Recall (%)	F Measure (%)	ROC (%)
Cross Validation	SVM	47.8	58.0	52.4	69.7
	k-NN	64.3	64.8	64.5	72.7
	RF	68.4	71.2	69.4	86.4
Percentage Split	SVM	47.8	56.9	51.9	69.3
	k-NN	68.0	67.8	67.8	74.4
	RF	71.4	73.4	72.1	87.8

The most successful classification result of white wine sample qualities after applying PCA was obtained by using Random Forest algorithm for both test modes. The accuracy of each cross-validation and percentage split mode with this algorithm is 69.9061%, and 67.449% respectively.

The most successful classification result of red wine sample qualities after applying PCA was obtained by using Random Forest algorithm for both test modes. The accuracy of each cross-validation and percentage split mode with this algorithm is 71.232%, and 73.4375% respectively.

4. Conclusions

For each classification model, we analysed how the results vary whenever test mode is changed. The study includes the analysis of classifiers on both red and white wine data set. The results are described in percentage of correctly classified instances, precision, recall, F measure, and ROC after applying the cross-validation or percentage split mode.

Different classifiers like k-nearest-neighborhood, random forests, and support vector machines are evaluated on datasets.

Results from the experiments lead us to conclude that Random Forests Algorithm performs better in classification task as compared against the support vector machine, and k-nearest neighbourhood.

After applying PCA, the success rate of quality classification for white wine has decreased from 70.3757% to 69.9061% for cross validation mode. The success rate of quality classification for white wine has decreased from 68.6735% to 67.449% for percentage split mode.

After applying PCA, the success rate of quality classification for red wine has increased from 69.606% to 71.232% for cross validation mode. The success rate of quality classification for red wine samples has increased from 71.875% to 73.4375% for percentage split mode.

References

[1] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, "Modelling wine preferences by data mining from physicochemical properties," *In Decision Support Systems, Elsevier*, 47 (4): 547-553. ISSN: 0167-9236.

[2] S. Ebeler, "Linking Flavour Chemistry to Sensory Analysis of Wine," in *Flavor Chemistry*, Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409-422.

[3] V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," in *Electronic Noses and Tongues in Food Science*, Cambridge, MA, USA: Academic Press, 2016, pp. 137-151.

[4] A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] S. Kallithraka, IS. Arvanitoyannis, P. Kefalas, A. El-Zajouli, E. Soufleros, and E. Psarra, "Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin," *Food Chemistry*, 73(4): 501-514, 2001.

[6] N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer," *Instrum. Measurement, IEEE Trans.*, 57: 2421-2436, 2008.

[7] S. Shanmuganathan, P. Sallis, and A. Narayanan, "Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality," *IEEE International Conference on Computational Intelligence Communication Systems and Networks*, pp. 82-89, July 2010.

[8] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," *IEEE International Conference on Data Mining Workshop*, pp. 142-149, Dec. 2014.

[9] UCI Machine Learning Repository, Wine quality data set, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[10] J. Han, M. Kamber, and J. Pei, "Classification: Basic Concepts," in *Data Mining Concepts and Techniques*, 3rd ed., Waltham, MA, USA: Morgan Kaufmann, 2012, pp. 327-393.

[11] J. Han, M. Kamber, and J. Pei, "Classification: Advanced Methods," in *Data Mining Concepts and Techniques*, 3rd ed., Waltham, MA, USA: Morgan Kaufmann, 2012, pp. 393-443.

[12] W. L. Martinez, A. R. Martinez, "Supervised Learning" in *Computational Statistics Handbook with MATLAB*, 2nd ed., Boca Raton, FL, USA: Chapman & Hall/CRC, 2007, pp. 363-431.

 This journal is © Advanced Technology & Science 2013