Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 305–312

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Selection of important features and predicting wine quality using machine learning techniques

Yogesh Gupta*

GLA University, Mathura, Uttar Pradesh, INDIA

Abstract

Nowadays, industries are using product quality certifications to promote their products. This is a time taking process and requires the assessment given by human experts which makes this process very expensive. This paper explores the usage of machine learning techniques such as linear regression, neural network and support vector machine for product quality in two ways. Firstly, determine the dependency of target variable on independent variables and secondly, predicting the value of target variable. In this paper, linear regression is used to determine the dependency of target variable on independent variables. On the basis of computed dependency, important variables are selected those make significant impact on dependent variable. Further, neural network and support vector machine are used to predict the values of dependent variable. All the experiments are performed on Red Wine and White Wine datasets. This paper proves that the better prediction can be made if selected features (variables) are being considered rather than considering all the features.

© 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications.

Keywords: Linear regression; neural network; support vector machine; wine quality

1. Introduction

Today, all type of industries is improving by adopting new technologies and applying these in all areas. These technologies are also helpful to enhance the production and making the whole process smooth. But, still there are different areas, which demands human expertise such as product quality assurance. Nowadays, it becomes an

---

* Yogesh Gupta. Tel.: +0-91-875-504-5936;
E-mail address: yogesh.gupta@gla.ac.in; er.yogeshgupta@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.041

expensive process as the demand of product is growing over the time. Therefore, this paper explores different machine learning techniques such as linear regression, neural networks (*NN*) and support vector machines (*SVM*) for product quality assurance. These techniques performs quality assurance process with the help of available characteristics of product and automate the process by minimizing human interfere. The work also identifies the important features to predict the values of dependent variables.

In this work, all above mentioned machine learning techniques are used to support wine industry. Wine quality assessment is one of the key elements in this context and this assessment can be used for certification. Such type of quality certification helps to assure wine quality in market. Wine has various characteristics like density, pH value, alcohol and other acids. Wine quality can be assessed by two types of tests; first is physicochemical test and second is sensory test [1]. Physicochemical test can be determined by lab tests and no human expert is required but for sensory test, a human expert is required. Moreover, Wine quality assessment is very difficult as the relationships between the physicochemical and sensory analysis are complex and still not fully understood [2].

In literature, some researchers have used machine learning techniques to assess wine quality, but still a huge scope is available for improvement. Sun et al. [3] predicted *six* geographic wine origins based on neural networks fed with *15* input variables. They used 170 samples of data from Germany for their experiments. They got *100*% predictive rate. Vlassides et al. [4] also used neural network for classification of Californian wine. Grape maturity level and chemical analysis are used for wine classification. A sample of *36* examples was used for experiments and achieved only *6%* error. Moreno et al. [5] classified 54 wine samples into two red wine classes using probabilistic neural network. Yu et al. [6] classified 147 bottles of rice wine to predict *three* categories of wine using spectral measurements. Beltran et al. [7] used SVM, neural network and linear discriminate analysis to classify Chilean wine. The experiments and analyses were performed on three different varieties of Chilean wine. Cortez et al. [8] compared several classification of wine dataset. Jambhulkar et al. [9] used various techniques to predict heart disease using wireless sensor network. They collected data from Cleveland dataset and extracted important attributes to predict heart disease. Zaveri et al. [10] predicted different diseases like TB, cancer, diabetes etc. using data mining techniques.

In this paper, linear regression, NN and SVM are implemented to determine dependency of wine quality on different *11* physicochemical characteristics. Moreover, the predictions are also made for wine quality on the basis of important variables/characteristics, selected according to their dependencies.

The paper is organized as follows: Section 2 provides the description and statistics of dataset used in this work. Section 3 discusses the proposed methodology in detail. Experimental results and analysis are explained in section 4. Conclusion is drawn in section 5.

2. Dataset

In this work, Wine dataset is used for all the experiments. Wine dataset is a collection of white and red wines [11]. White wine consists of *4898* samples and red wine contains *1599* samples. Each sample of both types of wine consists of *12* physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality rating. The quality rating is based on a sensory test carried out by at least three sommeliers and scaled in 11 quality classes from *0* - very bad to *10* - very excellent.

It is not possible to use both type of wine collections without preprocessing due to some deficiencies. One of the major deficiencies is the large amplitude of variable values e.g. sulfates (*0.3–2*) vs. sulfur dioxide (*1–72*). Moreover, some variables have values between 0 and 1. Such type of inconsistency may affect predictions due to more influence making by some variables than others. One of the ways to deal with such problem is linear transformation. Linear transformation can be achieved by dividing all the input values by maximum variable value.

3. Proposed methodology

In this work, machine learning techniques are used to determine dependency of wine quality on other variables and in wine quality predictions. This section gives insights of proposed methodology. First Wine dataset is pre-processed as explained in previous section. Further, linear regression is applied to determine dependency of Wine quality on other *11* independent variables (predictors). Then after, important predictors are selected according to dependency of wine quality on independent variables. At last, Wine quality is predicted with the help of support vector machine and neural network considering all predictors and selected predictors. The working of used machine learning techniques is discussed in following subsections.

3.1. Linear regression

Linear regression models the relationship between a dependent variable and two or more than two independent variables. It is used to predict the value of a variable based on the value of two or more other variables. It also allows determining the overall fit of the model and the relative contribution of each of the predictors to the dependent variable.

A linear regression model with *k* number of predictors (independent variables) and one response (dependent) variable can be expressed as (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \epsilon \tag{1}$$

where, Y is response variable and X_i are predictors (independent variables). ϵ is the residual term of the model, which is used for inference on the remaining model parameters. $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are regression coefficients.

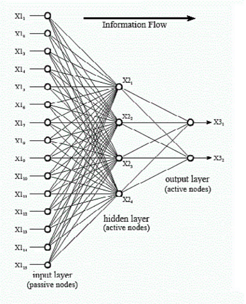

Fig. 1. Artificial neural network.

3.2. Neural network

Neural network is used by all human beings and animals to process information. It uses trillions of neurons for exchanging information through electrical pulses. But in the context of computer, NN is known as Artificial Neural Network (*ANN*). In this work, implemented neural network consists of *three* layers: input, hidden and summation, output as shown in Fig. 1. The function at input layer which passes input X_i to the hidden layer can be expressed as (2).

$$Y(X) = \frac{\sum_{i=1}^{n} Y_i \exp\left(\frac{-D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(\frac{-D_i^2}{2\sigma^2}\right)} \tag{2}$$

where σ is the width of the kernel. The output node computes *A/B*, which is *Y*.

The hidden layer consists of all training patterns X_i. When an unknown pattern X appears to this neural network then the squared distance between X and each X_i is computed. The squared distance can be expressed as (3).

$$Di^2 = (X - Xi)^T (X - Xi) \tag{3}$$

This squared distance is passed to kernel function. Then after, these distances are passed through summation function.

3.3. Support vector machine

SVMs are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. *SVM* constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. First, the model is trained by data, which is a set of points of the form as expressed in (4).

$$D = \{(x_i, c_i) \mid x_i \in \Re^p, c_i \in \{-1,1\}\}_{i=1}^{n} \tag{4}$$

where D is training dataset c_i represents the class which is either *1 or −1*. x_i indicates a data point which is a p-dimensional real vector and each x_i belongs to corresponding c_i.

SVM creates a $p-1$ dimensional hyperplane during training that separates data points into *two* classes. *SVM* selects *one* hyperplane which has maximal distance to the support vectors (nearest data points).

4. Experimental results and analysis

There are total *12* variables in both types of wine collections as discussed in section 2. The variable quality rating is considered as dependent variable and other *11* variables are assumed as predictors or independent variables in this work. Two types of analysis are done in this paper: firstly, the importance of each predictor for wine quality is identified and secondly, the value of wine quality is predicted using predictors.

Table 1. Linear regression summary of dependent variable (Quality) for Red Wine.

R= .60045958 R²= .36055170 Adjusted R²= .35611948 p < 0.0000

	Standard regression coefficient (b*)	Standard error of b*	Raw regression coefficient (b)	Standard error of b	t value	p-value
fixed acidity	0.053879	0.055944	0.0250	0.02595	0.96308	0.335653
volatile acidity	**0.240261**	**0.026851**	**1.0836**	**0.12110**	**8.94780**	**0.000000**
citric acid	-0.044038	0.035502	-0.1826	0.14718	-1.24044	0.214994
residual sugar	0.028513	0.026192	0.0163	0.01500	1.08860	0.276496
chlorides	**0.109230**	**0.024436**	**1.8742**	**0.41928**	**4.47007**	**0.000008**
free sulfur dioxide	**0.056491**	**0.028124**	**0.0044**	**0.00217**	**2.00864**	**0.044745**
total sulfur dioxide	**0.132979**	**0.029684**	**0.0033**	**0.00073**	**4.47983**	**0.000008**
density	-0.041789	0.050558	-17.8812	21.63310	-0.82657	0.408608
pH	**0.079080**	**0.036628**	**0.4137**	**0.19160**	**2.15897**	**0.031002**
sulphates	**0.192336**	**0.023999**	**0.9163**	**0.11434**	**8.01430**	**0.000000**
alcohol	**0.364470**	**0.034948**	**0.2762**	**0.02648**	**10.42901**	**0.000000**

4.1. Determining important features for prediction

Table 1 represents the regression summary for dependent variable i.e. quality for Red Wine. This table shows the dependency of quality on all predictors individually. Here, *R* is the co-relation coefficient which indicates how much dependent variable (quality) is co-related with all predictors as whole. R^2 and *adjusted* R^2 are also computed from *R*. The value of *Adjusted* R^2 is *0.3561* which shows *35.61%* dependency of quality on all predictors as whole. This value is less than *50%* which indicates that there are one or more predictors; those are not good for predicting the value of quality. At the same time, *P-value* is much less than 0.05 as shown in Table 1, which indicates that adjusted R^2 is significantly different from *Zero* and deny null hypothesis.

Table 1 also shows the standardized regression coefficients (*b**), standard error of *b**, raw regression coefficients (*b*), standard error of *b,* t-test (*t*) and *p-value*. The values of these coefficients are used to compare the relative contribution of each predictor in prediction of quality. The predictors such as volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates and alcohol are important predictors of quality as *p-value* for these predictors is less than 0.05 (95% confidence interval). The raw or un-standardized regression coefficient (*b*) for volatile acidity is 1.0836, which indicates that if all other predictors are controlled (constant) then increment of one unit in volatile acidity increases the quality by 1.0836. The same statement can be made for other predictors.

Similarly, Table 2 represents the regression summary for quality of White Wine. Table 2 clearly indicates that quality of White Wine is *28.02%* dependent on all predictors as whole. This value is less than 50%, it means there are one or more predictors; those are not good for predicting the white wine quality. Table 2 also shows that *p-value* for individual predictors such as fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates and alcohol is much less than 0.05, which means these predictors predict quality more significantly in comparison to others. The un-standardized regression coefficient (*b*) is 0.06 for fixed acidity, which indicates that if all other predictors are controlled (constant) then increment of one unit in fixed acidity increases the quality by 0.06. The same statement can be made for other predictors.

Table 2. Linear regression summary of dependent variable (Quality) for White Wine.

R= .53091465 | R^2= .30675369 | Adjusted R^2= .28025362 | $p < 0.0000$

	Standard regression coefficient (b*)	Standard error of b*	Raw regression coefficient (b)	Standard error of b	t value	p-value
fixed acidity	**0.062430**	**0.019889**	**0.066**	**0.02087**	**3.1389**	**0.001706**
volatile acidity	**0.212048**	**0.012951**	**1.863**	**0.11379**	**16.3733**	**0.000000**
citric acid	0.003019	0.013087	0.022	0.09577	0.2307	0.817589
residual sugar	**0.466653**	**0.043109**	**0.081**	**0.00753**	**10.8249**	**0.000000**
chlorides	-0.006100	0.013483	-0.247	0.54654	-0.4524	0.650973
free sulfur dioxide	**0.071681**	**0.016210**	**0.004**	**0.00084**	**4.4219**	**0.000010**
total sulfur dioxide	-0.013712	0.018142	0.000	0.00038	-0.7558	0.449791
density	**0.507528**	**0.064417**	**150.284**	**19.07451**	**7.8788**	**0.000000**
pH	**0.117021**	**0.017967**	**0.686**	**0.10538**	**6.5131**	**0.000000**
sulphates	**0.081374**	**0.012936**	**0.631**	**0.10039**	**6.2905**	**0.000000**
alcohol	**0.268840**	**0.033656**	**0.193**	**0.02422**	**7.9878**	**0.000000**

4.2. Predicting value of dependent variable (Quality)

Two different machine learning techniques neural network and *SVM* have been used to predict the wine quality in this work. The wine quality is predicted using all features of dataset and for selected features (determined from previous subsection) of dataset.

Table 3 represents neural network regression analysis for Red Wine data. In this work, 11-5-1 neural network architecture is used. Table 3 shows the predicted values against original values of quality for Red Wine in both the cases: for all features and for selected features. It is not feasible to show results for all documents, therefore the results are shown for few documents randomly. This table clearly indicates that wine quality is predicted more accurately for selected features in comparison to using all features. The overall summary of neural network regression for Red Wine dataset is also shown in Table 4. This table clearly depicts that the values of training error, testing error and validation error are less when selected features used for predictions rather than all features used for predictions.

Table 3. Predicting quality rating value using neural network regression analysis for Red Wine.

Document No.	Original Quality value	Quality Output (11-5-1) for all features	Quality Output (7-5-1) for selected features
1	5.000000	4.825105	5.023361
82	5.000000	5.118246	4.989526
228	5.000000	5.260592	5.116120
394	5.000000	4.780642	4.982879
709	6.000000	6.024094	6.007278
864	5.000000	4.966148	5.007411
1061	6.000000	6.267385	6.028679
1289	5.000000	5.210451	5.044965
1354	5.000000	5.209148	5.036600
1553	6.000000	5.935853	5.998063

Table 4. Overall summary of neural network regression analysis for Red Wine.

Network name	Training error	Test error	Validation error
(All features) MLP 11-5-1	0.187312	0.195660	0.169588
(Selected features) MLP 8-5-1	0.145736	0.146024	0.140383

Table 5 represents the predicting values of quality using SVM analysis for red wine dataset. This table proves that SVM gives more accurate predictions for selected features in comparison to using all features.

Table 5. Predicting quality rating value using SVM analysis for Red Wine.

Document No.	Original Quality value	Quality Output for all features	Quality Output for selected features
1	5.000000	4.571732	4.725478
82	5.000000	4.955524	4.995552
228	5.000000	5.246439	5.089935
709	6.000000	6.134489	6.097158
864	5.000000	4.707895	4.914136
1061	6.000000	5.279055	5.780928
1289	5.000000	4.716610	4.953137
1354	5.000000	5.393295	5.070797
1553	6.000000	5.839632	5.949534

Similar analysis is done for white wine as shown in Table 6-8. Table 6 represents neural network regression analysis for white wine data. In this work, 11-5-1 neural network architecture is used to predict the values. Table 6 shows the predicted values against original values of quality for White Wine dataset in both the cases: for all features and for selected features. This table clearly indicates that more accurate predicted values are obtained for selected features in comparison to all features. The overall summary of neural network regression for white wine data is also shown in Table 7. This table shows that the values of test error, training error and validation error obtained for selected features are less than the model which includes all features. Table 8 represents the predicting values of quality using *SVM* analysis for white wine dataset. This table proves that *SVM* gives more accurate predictions for selected features in comparison to using all features.

Table 6. Predicting quality rating value using neural network regression analysis for White Wine.

Document No.	Original Quality value	Quality Output (11-5-1) for all features	Quality Output (8-5-1) for selected features
1	6.000000	5.533817	5.796299
26	6.000000	5.774385	5.854882
82	6.000000	5.944356	5.998266
179	4.000000	4.592797	4.209509
657	6.000000	6.035119	6.014234
1179	5.000000	5.432416	4.997073
1828	6.000000	5.665227	5.852938
2665	7.000000	6.715398	6.803775
2736	6.000000	5.271117	5.617078
4732	6.000000	6.414085	6.205033

Table 7. Overall summary of neural network regression analysis for White Wine.

Network name	Training error	Test error	Validation error
(All features) MLP 11-5-1	0.234133	0.241568	0.243491
(Selected features) MLP 8-5-1	0.190578	0.207456	0.199758

Table 8. Predicting quality rating value using SVM analysis for White Wine.

Document No.	Original Quality value	Quality Output for all features	Quality Output for selected features
1	6.000000	5.694713	5.868677
26	6.000000	5.818790	5.951792
82	6.000000	5.076027	5.138848
179	4.000000	4.174945	4.094459
657	6.000000	5.848917	5.948470
1179	5.000000	4.585354	4.888971
1828	6.000000	5.650180	5.801910
2665	7.000000	7.429853	7.043530
2736	6.000000	5.797010	5.965150
4732	6.000000	6.087934	6.043868

Finally, it can be summarized that SVM is a better machine learning technique for wine quality predictions on the basis of results. At the same time, more precise predictions can be made by SVM and neural network using selected predictors rather than all predictors.

5. Conclusion and future directions

The interest has been increased in wine industry in recent years which demands growth in this industry. Therefore, companies are investing in new technologies to improve wine production and selling. In this direction, wine quality certification plays a very important role for both processes and it requires wine testing by human experts. This paper explores the usage of machine learning techniques in two ways. Firstly, how linear regression determines important features for prediction. Secondly, the usage of neural network and support vector machine in predicting the values. The benchmark Wine dataset is used for all experiments. This dataset has two parts: Red Wine and White Wine data. Red wine contains 1599 samples and white wine contains 4898 samples. Both red and white wine dataset consists of 12 physicochemical characteristics. One (quality) is dependent variable and other 11 are predictors. The experiments shows that the value of dependent variable can be predicted more accurately if only important features are considered in prediction rather than considering all features. In future, large dataset can be taken for experiments and other machine learning techniques may be explored for wine quality prediction.

References

[1] Ebeler S. (1999) “Flavor Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine”. *Kluwer Academic Publishers*, 409–422.

[2] Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) “Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception”. *Analytica Chimica Acta* **484 (1):** 33–34.

[3] Sun, Danzer and Thiel. (1997) “Classification of wine samples by means of artificial neural networks and discrimination analytical methods”. *Fresenius Journal of Analytical Chemistry* **359 (2)** 143–149.

[4] Vlassides, Ferrier and Block. (2001) “Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information”. *Biotechnology and Bioengineering* **73 (1)** 55-68.

[5] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) “Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks”. *Talanta* **72** 263–268.

[6] Yu, Lin, Xu, Ying, Li and Pan. (2008) “Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy”. *Agricultural and Food Chemistry* **56** 307–313.

[7] Beltran, Duarte-Mermoud, Soto Vicencio, Salah and Bustos. (2008) “Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer”. *IEEE Transactions on Instrumentation and Measurement* **57** 2421-2436.

[8] Cortez, Cerdeira, Almeida, Matos and Reis. (2009) “Modeling wine preferences by data mining from physicochemical properties”. *Decision Support Systems* **47** 547-553.

[9] Jambhulkar and Baporikar. (2015) “Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network”. *International Journal of Computer Science and Applications* **8 (1)** 55-59.

[10] Zaveri, and Joshi. (2017) “Comparative Study of Data Analysis Techniques in the domain of medicative care for Disease Predication”. *International Journal of Advanced Research in Computer Science* **8 (3)** 564-566.

[11] Portuguese Wine - Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt, July 2008.