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.

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
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