Wine Quality Estimator Based on Physiochemical Attributes

Ahmed Zarie
Department of Computer Science and Engineering
The American University in Cairo

Abstract

Proposed is a machine learning (data mining) approach to estimate red wine quality, based on easily collected analytical (physicochemical) characteristics of the wine. This paper is a comparative study of five techniques of classification; gradient boosting of decision trees (GB), logistic regression (LR), support vector machine (SVM) and an ensemble of all the classifiers with soft voting and hard voting. This model is to assist in wine evaluation on the production level.

Key words: SVM, Support Vector Machine, Logistic Regression, Gradient Boosting, Decision Trees, Data Mining.

1 Introduction

Advancement in information technologies has created the possibility to evaluate large amounts of data for the purpose of decision making. One of the subfields in this category is product quality assurance. There are several data mining models that used nowadays to tackle this particular point. There are familiar models used frequently; neural networks and SVM were used in particular for the wine quality problem, and this paper proposes two other techniques that show promising results; gradient boosting and ensemble soft voting.

When applying data mining techniques, variable selection is a critical point for consideration; kernel type for the SVM, layers architecture for the NN, and depth/estimators for the GB model. While variable selection is essential, and should be optimised for better results, it should also be tuned for performance constraints.

2 Data

The data used is gathered from the UCI dataset repository, *Wine Quality Dataset* [1]. It has 12 attributes per instance, with an overall of 1599 instance for red wine, which was used for this paper.

3 Validation

A k-fold cross validation approach has been taken to evaluate all models. With k=5, a total of 20 experiments were ran for each model (20 * 5 = 100 run) with the average mean absolute deviation (MAD) calculated for each model. The MAD per run is defined as:

$$MAD = \sum_{i=1}^{N} |y_i - \widehat{y}_i| / N$$

While the average MAD per model is the sum of all MADs over the total number of runs. Average MAD per model, will be referred to also as MAD

4 Used Models

1- SVM: The support vector machine model was tuned with an rbf kernel, with 1000 iterations as its upper bound.

2- Logistic Regression: tuned with inverse of regularisation strength to be 1.0

3- Gradient Boosting: build over decision trees -*estimators*- (1000 estimator), and a maximum depth for the tree = 10

4- Learning ensemble with soft voting: ensemble of SVM, LR, and GB - to take the average of the resulting probabilities for the final decision (soft voting).

5- Learning ensemble with hard voting: ensemble of SVM, LR and GB - to take the decision based on the class with the most votes.

5 Comparative Results

For the results, I will be comparing with a paper that uses the same dataset with same validation techniques. For comparison, I will be using the MAD. The paper is *Modeling wine preferences by data mining from physicochemical properties* [1]. The paper will be referred to by MWP, and this paper will be referred to WE.

	SVM	GB	LR	EN-S	EN-H	NN	MR
MWP	0.46	-	-	-	-	0.51	0.50
WE	0.46	**0.36**	0.46	0.38	0.45	-	-

SVM: Support Vector Machine

GB: Gradient Boosting

EN-S: Ensemble Soft Voting

EN-H: Ensemble Hard Voting

NN: Neural Network

MR: Multiple Regression

As per the demonstrated results, the GB approach yield the best results compared to the MWP's best performance (SVM), with MAD equal to 0.36 for the GB vs 0.46 for the SVM. Another approach that also shows better results than the SVM, is the soft voting approach with MAD = 0.38

6 Code

Code and data folders could be found via the following link:
github.com/zare3/WineQualityEstimator

7 References

[1] Cortez, Paulo et al. "Modeling Wine Preferences By Data Mining From Physicochemical Properties". *Decision Support Systems* 47.4 (2009): 547-553. Web.