Predicting Harbertson-Adams Assay Phenolic Parameters In Red Wines Using Visible Spectra

Anibal Catania[1], Carlos Catania[2], Santiago Sari[1], and Martín Fanzone[1]

[1] INTA - EEA Mendoza
catania.anibal|fanzone.martin|sari.santiago@inta.gob.ar
[2] UNCuyo - Facultad de Ingenieria - LABSIN
harpo@ingenieria.uncuyo.edu.ar

Abstract. The Harbertson-Adams phenolic parameter assay is a well-known method to measure a panel of phenolic compounds in red wines. However, the multistep analyses required by the method fail at producing results on multiple parameters rapidly. In the present article, we analyze the benefits of applying a statistical model based on Principal Component Analysis (PCA) and a statistical learning technique denoted as Support Vector Regression Machines (SVR) for correlating sample spectra data to the Harbertson-Adams assay, on each of the phenolics components. The resulting model showed a high correlation between the measured and predicted values for each of the phenolic parameters despite the multicollinearity and high dimensions of the dataset.

Keywords: Phenolic components · Wine making · Statistical learning

1 Introduction

The importance of phenolic compounds on sensory perception and quality of red wines requires that they should be readily quantified at all stages of winemaking. The Harbertson-Adams [2] phenolic parameter assay is a well-known method to measure a panel of phenolic compounds in red wines. However, the winemaking industry needs to assay thousands of samples with minimal compound errors and, the multistep analyses required by the method are sometimes impractical for producing results on multiple parameters rapidly. Moreover, in the particular case of the Argentinean winemaking industry, most wineries lack the required knowledge base for adequately conducting the analyses.

In the last ten years, a few approaches based on Spectroscopy predictive methods have emerged as a promising alternative to current analytical methods. The general idea behind predictive methods consists of the application of statistical learning techniques to correlate information from sample spectra to an analytical reference method. Once a model is built and validated, it is possible to obtain the composition of unknown samples from their spectra.

In 2008 Skogerson, et al [3]. conducted a study to investigate the application of the partial least squared (PLS) statistical method to UV-visible spectra of wine samples to develop predictive models for determining a range of phenolic

components at all stages of winemaking. The study carried out over 400 samples analyzed a wide range of grape cultivars and growing regions.

The collected spectra data suffer problems that can difficult the task of statistical learning models: a) the high correlation between different bands from the spectra (multicollinearity) and b) the high dimensionality of the data.

In the present article, we analyze the benefits of applying a combination of (PCA) Principal Component Analysis and a well-known statistical learning technique denoted as Support Vector Regression Machines (SVR) [1] for correlating sample spectra data to the Harbertson-Adams assay. The hypothesis is that the PCA/SVR capability of dealing with both multicollinearity and high dimensional data will produce a model with a more reliable prediction of phenolic parameters.

The main contributions of the present article are:

- A new dataset with 538 samples with information of the UV-visible spectra and the Harbeston-Adams assay for different Argentinean grape cultivars.
- A preliminary study of the viability of applying PCA/SVR for prediction of phenolic parameters.

2 Material and Methods

2.1 Dataset

All the wines were made at INTA Wine Research Center experimental winery, and they were from different vintages. The dataset includes 537 samples. The predominantly variety was Malbec (n=498) being the most cultivated grape in Argentina, but the dataset also includes wines from Bonarda (n=30) and from Cabernet Franc (n=9). The wines also came from different growing regions. The Malbec wines came from Ugarteche (n=18) and Drummond (n=24) in Luján de cuyo Deparment. In addition, some Malbec samples came from Cruz de Piedra (n=12) located in Maipú, and from San Pablo (n=444) located in Tupungato. All the Bonardas wines came from Lavalle, and Cabernet Franc wines came from Agrelo (Luján de Cuyo).

The sample was taken when the wine was bottled, and they were centrifuged prior to analysis in a CM4080 centrifuge (Rolco, Buenos Aires, Argentina). Absorption spectra of each wine sample were collected from 380–750 nm, at 1 nm interval, using a Perkin-Elmer Lambda 25 UV-visible spectrophotometer (Norwalk, CT, USA) with 0.1 mm path length flow cell. After that, total phenols (FT), Anthocyanins (AT), small polymeric pigments (SPP), large polymeric pigments (LPP), and total tannins (TT) were measured as described by (Harbertson et al. 2003)[2]

2.2 Experiment Setup

The applied methodology for predicting phenolic parameters consists of the application of the PCA statistical analysis for dimensionality reduction, followed by the application of the SVR regression algorithm for minimizing the generalization error on unseen examples.

The evaluation of the PCA/SVR was carried out following the usual statistical learning methodology. The dataset described in section 2.1 was random and uniformly split in an 80/20 ratio. The 80% of the dataset was used for calibrate the PCA/SVR parameters and build the model, whereas the remaining 20% (testing set) was used for evaluating the performance of the resulting PCA/SVR model on unseen examples. A different model was built for each of the phenolic components. Each model component was tuned after applying a 2x5 Cross Validation resampling technique on the 80% of the calibration dataset. The configuration of the resulting model is presented in Table 1 for reproducibility purposes. The third and fourth columns of the table refer to the SVR hyperparameters configuration (sigma and C), whereas the second column refers to the number of principal components (PC) used as predictor variables by SVR.

3 Preliminary Results

Notice that all the results shown in this section correspond to the prediction of the generated models on the 20% of the unseen samples. The columns fifth and sixth from table 1 show the results in terms of the root mean squared error (RMSE) and the coefficient of determination (R^2). As can be observed, the coefficient of determination (R^2) for all parameters was greater than 0.84, with the exception of large polymeric pigment and total tannins (both with values around 0.72%). It is possible to explain the similar behavior of the two parameters since they have the same prior step in the lab assays. On the other hand, the method seems to be not very accurate when wines present low tannin content.

Table 1. Selected parameters for the resulting models after 2x5 cross validation on the calibration dataset. The number of Principal Components (PC) selected as input variables of the SVR. The *Sigma* and *C* parameters of the SVR model

	PCA	**SVR**			
Phenolic Compoment	**number of PCs**	**Sigma**	**C**	R^2	**RSME**
AT	30	0.005	16	0.8754	59.763
SPP	30	0.0062	16	0.8493	0.1804
LPP	30	0.025	6	0.7250	0.2332
TT	30	0.25	13	0.7250	77.335
FT	30	0.025	13	0.8540	0.2332

In Fig. 1, the values for each of the Harbertson-Adams assay parameters were plotted against values predicted by the model for each parameter based on the UV visible spectrum of each sample. The figure confirms a strong correlation between the measured and predicted values for each of the Harbertson-Adams assays phenolic parameters.

4 Concluding remarks and Future work

The application of the PCA/SVR method for predicting Harbertson-Adams assay phenolic parameters exhibited encouraging results. The resulting models

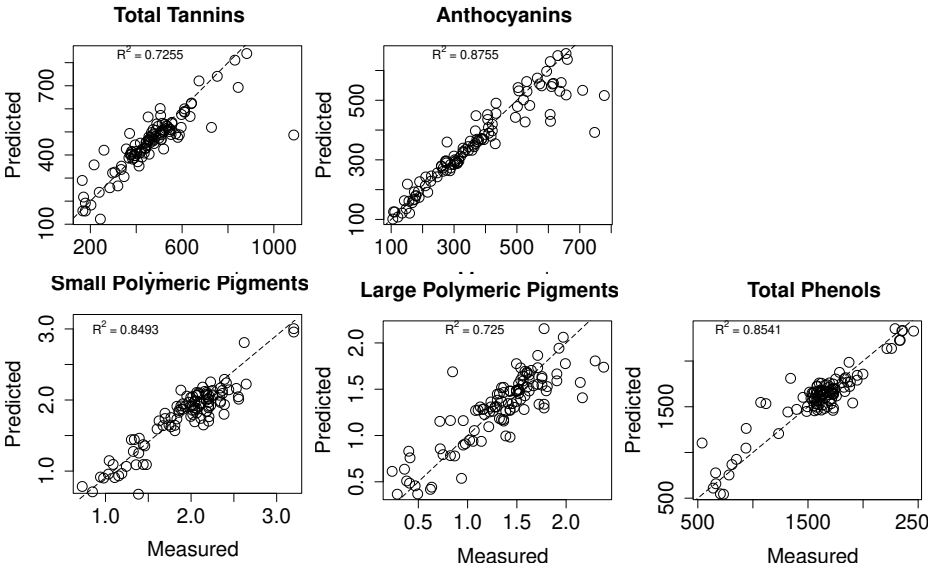

Fig. 1. Correlation between the measured Harbertson-Adams assay phenolic parameters and the values predicted by the models based from the UV-visible spectra for the test data set. The coefficient of determination (R^2) is included for each phenolic component.

showed a high correlation between the measured and predicted values for each of the phenolic parameters despite the multicollinearity and high dimensions of the dataset. However, in order to provide a simple method for producing rapid results, several more aspects require a more in-depth analysis, the viability of the method requires verification in more samples with higher variability (including different grapes cultivar and different photo spectrometers).

Finally, a comparison with the method proposed by Skogerson et al. [3] as well as the possibility of producing per-cultivar models are two possible aspects to consider in future works.

References

1. Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: Advances in neural information processing systems. pp. 155–161 (1997)
2. Harbertson, J.F., Picciotto, E.A., Adams, D.O.: Measurement of Polymeric Pigments in Grape Berry Extract sand Wines Using a Protein Precipitation Assay Combined with Bisulfite Bleaching. American Journal of Enology and Viticulture **54**(4), 301–306 (2003), https://www.ajevonline.org/content/54/4/301
3. Skogerson, K., Downey, M., Mazza, M., Boulton, R.: Rapid determination of phenolic components in red wines from uv-visible spectra and the method of partial least squares. American Journal of Enology and Viticulture **58**(3), 318–325 (2007), https://www.ajevonline.org/content/58/3/318