Cost efficient prediction of Cabernet Sauvignon wine quality

Răzvan Andonie
Computer Science Department
Central Washington University
Ellensburg, USA
and
Electronics and Computers Department
Transilvania University, Braşov, Romania
Email: andonie@cwu.edu

Anne M. Johansen
Department of Chemistry
Central Washington University
Ellensburg, USA
Email:johansea@cwu.edu

Amy L. Mumma
L'École de Business
Institut Américain Universitaire
Aix-en-Provence, France
Email: amy.mumma@iaufrance.org

Holly C. Pinkart
Department of Biological Sciences
Central Washington University
Ellensburg, USA
Email: pinkarth@cwu.edu

Szilárd Vajda
Computer Science Department
Central Washington University
Ellensburg, USA
Email: szilard.vajda@cwu.edu

***Abstract*—The quality of wines can be assessed both from chemical/biological tests and sensory tests (which rely mainly on human experts). Determining which is the subset of tests to be used is a difficult problem. Each test has its own contribution for predicting the quality of wines and, in addition, its own cost. We use our own database, consisting of 32 wine characteristics applied to 180 wine samples. In addition we use wine quality labels assigned by a wine expert. To the extent of our knowledge, this is the first study of this kind on wines from Washington State, and also the first wine study in general to include cost minimization of the measurements as a goal. Our approach is based on two stages. First, we identify reasonably good classifiers (from a given set of classifiers). Next, we search for the optimal subset of features to maximize the performance of the best classifier and also minimize the overall cost of the measurements. As a result, through our method we can answer queries like "the best performing subset of tests for a given threshold cost".**

I. INTRODUCTION

Home to Microsoft, Boeing, Starbucks, Costco, Expedia, and Amazon, Washington State is also a premium wine producing region that is (after California) the second largest producer in the USA, with more than 50,000 acres (20,234 hectares) of vines and more than 850 wineries (Washington State Wine Commission[1], 2016). The wine industry has become a respected and influential $4.4 billion-plus business within Washington State. Washington wine is available in 50 states and more than 40 countries globally. Washington State has a diversity of landscapes, from evergreen coasts and snow-capped mountains to a sagebrush desert, consisting of 13 unique growing regions. More than 40 grape varieties are cultivated, including Riesling, Chardonnay, Cabernet Sauvignon, Merlot and Syrah.

Cabernet Sauvignon is a variety grown around the world in diverse soils and climates, each of which produces very different styles of wine. This varietal is characterized by thick skin, moderate to high acidity, full body, strong tannins and an affinity for oak aging. Flavor characteristics include black and red fruits, blackberry, black currant, plum, black cherries, mint, and tobacco among many others.

Wine certification and quality assessment are key elements within the wine industry. Quality evaluation is often part of the certification process and can be used to improve wine making by identifying the most influential factors. The quality of wines can be assessed both from chemical/biological tests (determination of alcohols, acidity, reduced sulfur, bacteria, etc.) and sensory tests (which rely mainly on human experts). However, since taste is the least understood of the human senses [1] and no clear relationship exists with biochemical markers [2], sensory testing remains difficult to formalize. Here, we attempt to finding the most cost efficient set of biochemical markers that allow for optimal characterization of wine quality as determined by an expert.

Finding which is the subset of tests to be used is a difficult problem. Each test has its own contribution for predicting the quality of wines and, in addition, its own cost. Therefore, choosing the optimal subset of tests from a set of biochemical tests is a multi-criterial hard optimization problem. Meanwhile, using wine experts to assess the quality of wines has its own drawbacks: subjective judgments, expensive, lack

[1] http://www.washingtonwine.org/

of experts, etc. It would be highly beneficial to the wine producers to create a set of tests to automatically predict the quality of wines.

Smaller wineries often lack access to biochemical testing and analysis technology for detection of faults. This may result in entire lots of wine affected by one or more faults, creating a lower quality wine. In some cases, the quality is so severely impacted that the wine cannot be sold. This can clearly impact the financial state of the individual winery, but will also have an impact on its contribution to the local, state, and national economy. It is therefore important to understand the nature of wine faults and to be able to classify efficiently wines according to their quality, in particular when faults are imminent at the low quality end.

In real-world applications, we have a large variety of wines, chemical/biological tests, and machine learning tools. At some point, each wine producer chooses the right combination of these in order to optimize the decision process. Almost all previous data mining studies on wines were performed on relatively small datasets (both in terms of number of wines and number of features) because obtaining such datasets is expensive, a dataset cannot be re-used for "similar" wines, and labeling the wines by human experts is also costly.

Each measurement is associated with a cost that can vary widely from one test to another. This makes such comprehensive testing prohibitively expensive for routine work. Therefore, only small datasets exist that allow for a relationship analyses between biochemical characteristics and quality of wine.

We seek to tackle this issue with two layers of optimization by finding *a)* the "best" classifier from a family of models and *b)* the optimal subset of features to maximize the performance of this classifier, while minimizing the overall cost of analysis.

In our study, we use 60 randomly chosen Washington State Cabernet Sauvignon wines for analysis and three bottles of each wine are tested, resulting in a total of 180 wine samples. A series of 30 biochemical tests are performed on each wine. Two additional features (age and region) are added to the input features. Each wine is labeled by a human expert into six categories, according to its overall quality. Our strategy is to train a statistical classifier - considering supervised training, to predict the wine quality for unseen wines.

To the extent of our knowledge, this is the first study of this kind on wines from this region and also the first wine study in general to include cost minimization as a goal. Compared to similar studies, our dataset is large (more wines and more biochemical tests). We use several classifiers and feature selection tools and try to find an optimal subset of tests which could be used to simultaneously maximize the prediction performance of wine quality and minimize the overall cost of the measurements.

The remaining part of the paper is structured as follows. Section II presents similar approaches reported in the literature. Section III describes the dataset and the attributes of the 180 wine samples. Section IV introduces the wine quality prediction method. Section V presents experimental results with their interpretation. Section VI concludes with some final remarks.

II. Related work

In wine technology, machine learning can be used for classification of wines according to origin, producers, type, and for optimization of wine blending and electronic nose for sensory analyses. Neural classifiers and discriminant techniques have been used to classify, verify the wine origin, or predict properties of: Chilean wines [3], Slovak wines [4], [5], Montelpuciano d'Abruzzo Italian wines [6], Canary Islands' wines [7], Romanian wines [8], Spanish wines [9], [10], [11], Portuguese wines [11], and Italian wines from Atripalda [12]. Neural-network-assisted optimization of wine blending was proposed in [13], [14].

A. Methods

An overview of neural network applications in wine technology can be found in [15]. A variety of machine learning methods have been used: multilayer perceptron (MLP) using quick-propagation and quasi-Newton propagation training [4], [8], MLP using backpropagation [11], [12], [14], time delay neural networks trained with the MLP Levenberg-Marquadt method [16], linear discriminant analysis [6], neuro-fuzzy models [12], MLP trained with the Broyden-Fletcher-Goldfarb-Shano learning rule [13], stepwise linear discriminant analysis and neural networks [9], decision trees [17], [18], principal component analysis (PCA) and cluster analysis (k-means) [19], probabilistic neural networks [8], least-squares support vector machines (SVMs) [20], [21], Kohonen self-organizing maps [7], etc.

From a machine learning perspective, a complex study can be found in [10]. The objective of that study was to find a classification model able to precisely differentiate between existing grape varieties (i.e., assuring the authenticity), and to assess the discriminatory power of different family compounds over the following classifiers: SVMs, random forests, MLPs, k-neighrest neighborks, and naïve Bayes. Given the fact that PCA was not able to accurately separate all the wine varieties, the best classification accuracy was obtained by the random forest algorithm. Although the random forest algorithm was able to perfectly classify all the grape origins, this was only possible when using all input features. The MLP classifier was the most accurate algorithm when dealing with less information.

B. Datasets

Machine learning depends largely on the datasets used for training. The well-known Wine dataset from the UCI repository[2] which contains 178 examples, with measurements of 13 chemical constituents, is less relevant in our case since it is very easy to discriminate and has been mainly used as a benchmark for new classifiers. Collecting data from wine samples is generally a tremendous logistic problem and financial burden is associated with the analytical measurements. This

[2] https://archive.ics.uci.edu/ml/datasets/Wine

explains why existing published datasets are generally small (with the exception of the dataset introduced in [22]):

- Urtubia *et al.* [19] performed the chemical analysis of 29 compounds taken around-the-clock, which produced around 22,000 measurement data.
- Chandra *et al.* [17] used 178 wine samples with 13 chemical descriptors (features).
- Kruzlikova *et al.* [4] used 36 samples characterized by 19 attributes. They selected the most relevant seven features for their classifier.
- Chichelli *et al.* obtained their data from the chemical analysis of 116 wine samples [6].
- Aires-De Sousa *et al.* [11] analyzed 21 samples on the basis of 15 anthocyanin[3] contents.
- Magarino *et al.* measured 19 features in 70 wines [9]. The effect of each input feature on the wine classification was evaluated in the form of the causal index, calculated from the trained neural networks.
- Bednarova *et al.* [5] conducted analyses for 739 wine samples to determine 11 enological measurements used as input features for their neural classifier.
- Gomez *et al.* [10] identified and quantified 41 volatile compounds of a total of only 42 wine samples.
- Hosu *et al.* [8] analyzed 81 wine samples using six analytically measured characteristics.
- In [21], Yu *et al.* used spectral measurements from 147 bottles to predict three categories of rice wine age.
- Yamazaki *et al.* generated their dataset automatically [16]. Sensors connected to a computer where exposed to the samples. This explains why this database is in the order of thousands.

The wine quality dataset, created by Cortez *et al.* and available from the UCI repository[4], is significantly larger and has been used by several authors as a benchmark for wine quality prediction methods [22], [18], [23]. It contains 4898 wine samples related to red and white variants of the Portuguese *vinho verde* wine, each with 11 physico-chemical attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) and one sensory data attribute (a quality score).

Practically, each study uses a different set of measurements. This is also because the contemporary global wine industry is inherently geographical, the origins of the grapes being a predominant factor in the promotion of the product. Typically, researchers in the aforementioned studies pick the physico-chemical tests to be performed and, in some cases, look for the optimal subset of tests (feature selection), using the performance of the classifier as a fitness function. In addition, in some studies, optimal classifiers were sought for a particular application, from a list of considered classifiers.

C. *Wine quality prediction*

Wine quality can be assessed by biochemical measurements. For instance, the Vis-NIR spectrometer system was suitable for wine quality determination [24]. In contrast to this, we focus here on wine quality assessed by human experts and we attempt to predict human wine taste preference by relating it to a wide range of biochemical analytical techniques. There are only few reported results on this topic [22], [18], [23].

In [22], each wine from the wine quality dataset was graded by experts, on a scale from 0 to 10. The authors adopted a regression approach, which preserves the order of the preferences: if the true grade is 3, then a model that predicts 4 is better than one that predicts 7. The regression model was trained in a supervised way to predict wine quality. The SVM regression with Gaussian kernel proved to provide the prediction best performance. Sensitivity analysis (as a fitness criterion) combined with the backward selection method were used for feature dimension reduction. The variable and model selection is performed simultaneously: in each backward iteration several SVM hyperparameters are searched, with the one that presents the best generalization estimate selected. The output of the SVM regression model is mapped to the nearest class (wine quality grade). The overall precision (positive predictive value) obtained was 62.4% for the red wine, respectively 64.6% for the white wine.

Appalasamy *et al.* tried unsuccessfully to improve the results from [22] using information gain attribute evaluation for feature selection and two classifiers (ID3 and naïve Bayes) [18]. On the same dataset, Nachev *et al.* [23] tested four predictive models: SVMs, cascade-correlation neural networks, general regression neural networks, and the MLP. The SVMs with polynomial kernel outperformed the three neural network models. They found that the best feature selection techniques are Chi-squared attribute evaluation and symmetrical uncertainty ranking.

III. Analyses Performed on Washington Cabernet Sauvignons

The dataset used in our investigation stems from a study performed in 2009 to determine the most common faults in Washington State Cabernet Sauvignon[5] wines.

Three bottles of 60 randomly chosen wines were tested to take into account bottle variation for a total of $n = 180$ bottles (wine samples). Organoleptic (appearance, aroma, palate), chemical and biological analyses were performed.

Each bottle was opened and the wine dispensed in a glove box, under nitrogen (N_2) atmosphere at 4°C, into three sets of sterile, amber glass vials. The vials were filled, leaving minimal head space, and sealed while inside the anaerobic chamber. This protocol was developed to prevent both chemical and photo-oxidation and to prevent loss of volatile organic compounds prior to analysis. Vials one, two and three

[3] Anthocyanins are water-soluble vacuolar pigments that may appear red, purple, or blue depending on the pH.

[4] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[5] In the United States, law requires that the grape varietal stated on the wine label constitute a minimum 85% of that varietal in the wine. The other 15% can be made up of any varietals the winemaker chooses, and this information does not have to be disclosed.

were analyzed organoleptically, chemically, and biologically, respectively.

For the chemical analyses, vials were handled by the CWU Chemistry Department, with most analyses outsourced to ETS Laboratories located in St. Helena, California (marked with an asterisk in Section A of Table I). Samples sent to ETS were kept chilled and shipped overnight on the date they were dispensed. At ETS they were analyzed within 7 days of dispensing. Chemical analyses at ETS were performed with standard techniques, including Fourier Transform Infrared Spectroscopy (FTIR, for pH and titratable acidity), Gas Chromatography with Flame Ionization Detection (GC-FID, for alcohols, acetaldehyde, ethyl acetate), Sequential Analyzer with UV-Vis Spectrophotometry (for volatile acidity), Aeration/Oxidation followed by Titration (for all forms of sulfur dioxide), and Gas Chromatography with Sulfur Chemiluminescence Detection (GC-SCD, for reduced sulfur compounds) [25]. At Central Washington University, in parallel, UV-Vis Spectrophotometry was used to carry out the optical analyses of compounds contributing to pigmentation (phenols, anthocyanins, etc) as well as GC-FID for alcohols and a pH meter for pH [26], [27], [25]. Section A of Table I shows number of observations made for each feature that was above the detection limit of the method (in parentheses). The following chemical features, with observations for all wine samples below detection limit (BDL) are considered as not relevant and will be not used: Hydrogen Sulfide, Diethyl Sulfide (DES), Methyl Mercaptan, thus leaving 25 chemical features that could be used for this study.

For the biological testing of bacteria and yeasts known to cause faults, two methods were used at CWU: (i) cultivation and identification of viable organisms and (ii) detection of DNA sequences of organisms from extracted DNA. For cultivation of organisms, portions of each wine sample were used within an hour of dispensing to inoculate (in triplicate) selective cultivation media specific of lactic acid bacteria (LAB) [28], acetic acid bacteria (AAB) [29], and *Brettanomyces bruxellensis* [30]. The cultured wine was incubated at 22-25°C for a minimum of 3 days for LAB and AAB, and up to 3 weeks for the *Brettanomyces* cultures. Following the appropriate incubation period, the average number of cultured organisms (reported as colony forming units, CFU) were determined for each wine. For the DNA amplification, aliquots from each wine were frozen immediately after dispensing and thawed just prior to DNA extraction. DNA was extracted from each sample using commercially available DNA extraction kits (MoBio Laboratories, Inc., Carlsbad, CA) designed for microbial DNA extraction from matrices containing organic material like humic acids, phenolics, and tannins. Extracted DNA was then subject to analysis via real-time polymerase chain reaction (PCR) [31], [32] to detect and quantify organisms. Section B of Table I lists features and number of observations that were above the detection limit of the method (in parentheses). There are 5 biological features.

In summary, 30 chemical/biological features were measured above the detection limit for at least some of the wines, with two parameters, DDEDS and AAB, showing particularly low frequency of occurrences of $n = 4$ and 9, respectively. These 30 chemical/biological features, each with an associated cost-per-analysis, were used in the present study to analyze their relationship to the overall organoleptic quality of the wine. Table I shows also the cost-per-analysis for each feature. In addition, wine age and wine region were included as features as they deemed to have significant discriminative power. Hence, we used $30+2 = 32$ features for our generated dataset.

For the organoleptic testing, wines were equilibrated to laboratory temperature (21°C) [33] to assure constancy throughout the analysis process. Wines were tasted on the same day of opening and standard tasting wine glasses were used. One of the authors of this paper is an expert taster[6] who rated each of the 180 sample wines based on appearance, aroma and taste on the following parameters: clarity, haziness, ropiness, color, oxidation, volatile acidity, ethyl acetate, sulfur dioxide, hydrogen sulfide, bacterial, maderization, oxidation. The scale ranged from 1 to 6, with wines scoring 1 being the least faulty in that particular parameter, and those scoring 6 as the most faulty.

These data were then used to generate an overall numerical score for each wine. The score spans from 1 to 6, 1 represents the best quality, while 6 designates a lower quality in this hierarchy. The score is used as a class label of the wine sample.

IV. Experimental Setup

With 180 samples and 32 input features, our dataset is large, both in terms of number of samples and biochemical markers (features), in comparison to other previous research attempts (see Section. II). The wine quality dataset in [22] is much larger than ours, but it only uses 11 features.

A. Methodology

We look for an optimal classifier for wine quality prediction which also minimizes cost. Our solution is based on two stages of optimization. First, we search for reasonably good classifiers (from a given set of classifiers). Next, we search for the optimal subset of features to maximize the performance of the best classifier and also minimize the overall cost of the measurements.

The labels were assigned on a subjective basis, therefore there is not an equidistant split among the different qualities. Based on this fact, even though a regression can be applied (as in [22]), such operation does not have a real meaning. Hence, our focus is shifted toward statistical classifiers such as decision trees, neural networks, nearest neighbor classifier, etc. as we would like to classify the quality of wines completely automatically, without involving a wine expert.

Considering the nature of our problem, we did not develop our own classifier set, but rather considered the well-known

[6] Amy Mumma, Professional Wine Woman 2005-2006, award received in Paris, 2007.

A. Chemical Analysis (25 features)		
Acids, Aldehydes and Alcohols* ($n = 180$)	Sulfur Containing Compounds* ($n = 180$)	Colorimetric Characteristics ($n = 180$)
pH, $10	Dymethyl Sulfide (DMS), $30	Pigmentation, Red (9 samples BDL), $2
Volatile Acidity, $26	Dymethyl Disulfide (DMDS) (80 samples BLD), $30	Pigmentation, Brown (9 samples BDL), $2
Titratable Acidity, $18	Diethyl Disulfide (DDEDS) (176 samples BDL), $30	Color Intensity (9 samples BLD), $2
Acetaldehyde, $110	Ethyl Mercaptan (107 samples BDL), $30	Copigmentation, $2
Ethyl Acetate, $110	Total Sulfur Dioxide (Total SO_2), $20	Color Anthocyanins, $2
Methanol, $100	Free Sulfur Dioxide (Free SO_2) (SO_2+ H_2O+HSO_3), $20	Polymeric Anthocyanins $2
1-Propanol, $25	Molecular Sulfur Dioxide (Molecular SO_2) (SO_2+SO_2+H_2O), $36	Total Phenols, $2
Isobutanol, $25		
2-Methylbutanol (Iso amyl alcohol), $25		
3-Methylbutanol (Active amyl alcohol), $25		
Ethanol, $25		
B. Biological Characteristics (5 features)		
Bacteria (by culture) ($n = 180$)	Bacteria (by DNA detection) ($n = 180$)	Fungi (by culture) ($n = 180$)
Lactic Acid Bacteria (11 samples BDL), $5	*Lactobacillus sp.* (42 samples BDL), $16	*Brettanomyces bruxellenis* (151 samples BDL), $5
Acetic Acid Bacteria (171 samples BDL), $5	*Pediococcus sp.* (52 samples BDL), $16	

TABLE I
Biochemical analyses used as input features. Number of total observations per feature (n), number of observations below detection limit (BDL), and approximate price-per-analysis are noted. *Analyses performed at ETS Laboratories, St. Helena, CA.

Weka[7] [34] framework, a collection of machine learning tools for different data mining tasks.

Our main performance criterion is the accuracy (the percentage of correctly classified instances). We also use the classification accuracy for individual classes. In all our experiments with classifiers and feature ranking tools we use 10-fold crossvalidation.

B. Choosing the classifier

We only consider the classifiers in Weka, and this allows us to use the Weka testing environment. Other classifiers may be also considered, but (we believe) with small chances to significantly improve the prediction accuracy. Using the default hyperparameters for each method, the best performing classifiers are listed in Table II. We have to note that the SVM classifiers performed very poorly on our data (the polynomial kernel SVM achieved only 40.00% accuracy), in contrast to the results reported in [22] and [23]. This confirms once again that there is no universal “best” classifier.

TABLE II
Performances of the most accurate Weka classifiers

Classifier	Accuracy
Random Forest	74.44%
IBk (1-nearest neighbor)	70.00%
Multilayer Perceptron	62.22%
KStar	68.88%
Random Committee	66.66%

We further tried to optimize the parameters of these classifiers, but we failed, meaning that the Weka implementations already resulted in reasonable performance. This is in accordance with the results of other authors [35].

As a result, we choose to use in the following the random forest classifier with its default Weka parameters: unlimited *maxDepth*, *batchSize* = 100, *numDecimalPlaces* = 2, *numFeatures* = 0, *numTrees* = 100, *seed* = 1 (see [36]).

C. Choosing the feature ranking method

We tested several feature ranking methods from Weka, including: *ChiSquaredAttributeEval*, *SymmetricalUncertAttributeEval*, *ReliefAttributeEval*, *InfoGainAttributeEval*, *CorrelationAttributeEval*. The methods based on information gain or on correlation between input features and classes perform best, among them *ChiSquaredAttributeEval* and *InfoGainAttributeEval*. Using these two feature ranking methods and the random forest classifier, we eliminated one-by-one the least important feature, showing each time the accuracy of the system with the reduced number of features. The results are shown in Fig. 1. Is to be observed, that even if the original 32 dimensional feature set is reduced by the 20 less important features provides still a reasonable good accuracy, which points us into the direction that the cost also can be reduced and still keeping high the prediction performances.

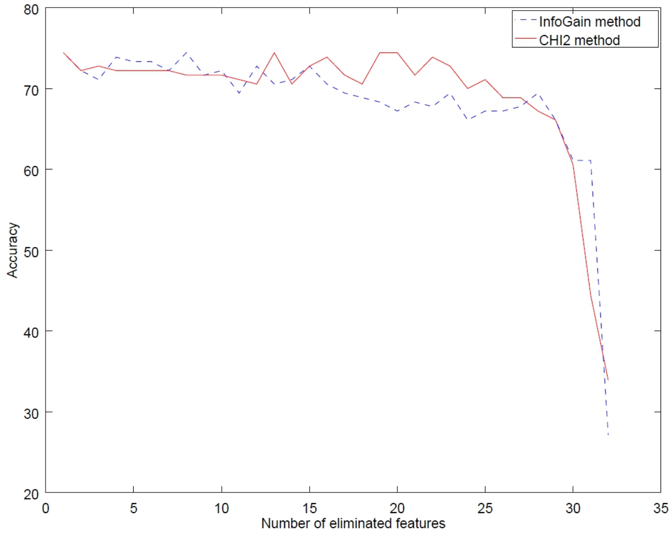

Fig. 1. Accuracy (in percent) vs number of eliminated features

Since we observe a slightly better performance with the Chi-squared feature ranking method [37], we choose to continue

[7] www.cs.waikato.ac.nz/ml/weka

with this method in combination with the random forest classifier, using for both their Weka implementations.

D. Cost optimization

Beside the accuracy, which is the primary goal of this research, we also consider the cost of each biochemical measurement, which ranges between \$2 to \$110.

Standard feature selection methods do not cope with the measurement cost issue. Cost-based feature selection was considered by few authors [38], [39], [40], but in completely different application areas than wine data mining. Including the measurement cost into the feature selection process generates a multi-criterial optimization problem (accuracy and cost) which is by its nature exponential, since all subgroups of features should be considered.

We scale to [0, 1] the average ranks generated by *ChiSquaredAttributeEval* (the smaller the rank value, the more important the feature is). Let r_i be the resulted value which corresponds to feature i. We also scale to [0, 1] the costs associated to the measurements (features). Let c_i be the scaled cost of feature i.

We define the new feature rank $new\_r_i = r_i * c_i$, for $i = 1, \ldots, 32$. This criterion favors cheap features with good discriminative power. Such features will have small new_r_i values. Other formulas may be also considered for computing new_r.

V. Results and discussion

Based on the method described above, the ranking (from most important feature to less important) is: co-pigmentation, age, region, red color, total SO_2, isobutanol, color intensity, volatile acidity, polymeric anthocyanins, total phenols, color anthocyanins, brown color, free SO_2, brettanomyces, acetic acid bacteria, pH, lactic acid bacteria, titratable acidity, ethanol, active amyl alcohol, iso amyl alcohol, pediococcus sp., 1-propanol, lactobacillus sp., ethyl acetate, molecular SO_2, ethyl mercantile, dymethyl disulfide, dymethyl sulfide, diethyl disulfide, methanol and acetaldehyde.

Using the computed feature ranks new_r_i, for $i = 1, \ldots, 32$, we eliminate one-by-one the least important feature (the feature with maximal new_r_i from the existing subset of features). This backward stepwise selection is not guaranteed to give us the best model containing a particular subset of features, but this is the price we pay in order to reduce the computational complexity from exponential to linear and also to avoid possible overfitting. Just because the best subset has a better model on the training data does not necessarily mean that it is really going to be a better model overall in the context of test data.

The result of this greedy optimization is shown in Fig. 2. From these results, the user (the chemist) can extract data (see Table III) like “the best performing subset of tests for a given threshold cost”.

For instance, spending \$198, we obtain an overall prediction accuracy of 71.11% using the following 20 features (in decreasing order of importance): co-pigmentation, age, region, red color, total SO_2, isobutanol, color intensity, volatile acidity, polymeric anthocyanins, total phenols, color anthocyanins, brown color, free SO_2, brettanomyces, acetic acid bacteria, pH, lactic acid bacteria, titratable acidity, ethanol, active amyl alcohol. Compared to the full panel of analyses, this represents a substantial cost reduction of 381% at the expense of only 4.5% loss in overall prediction accuracy.

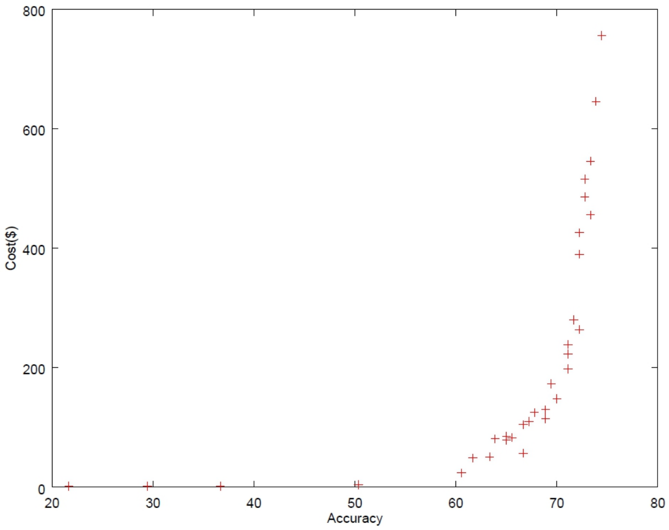

Fig. 2. Total cost vs accuracy (in percent)

Most noteworthy is that when we consider only the five most important features (i.e., co-pigmentation, age, region, red color, total SO_2), the accuracy drops by only 13.9% absolute reduction to 60.6%, but at a minimal cost of \$24, which is only 3.2% of the \$756 when all the features are considered.

A significant 10% overall increase in accuracy (66.7%) is obtained when adding three more features (8 features in total), but at the expense of a three fold increase in cost (\$77) and only 3% gain in accuracy of predicting the best wines.

To investigate wine category specific accuracies (see details in Section III), category 6 wines ($n = 8$) - which stands for the worst quality among the possible 6 categories investigated, were merged with category 5 wines ($n = 26$) -which is also a lower quality, resulting in 5 statistically more representative wine quality categories that ranged in sample size between $n = 32$ and 39. Interestingly, across all model runs with varying features, the best category specific prediction accuracies were observed for the ”best” (category 1) and ”least favorable” (merged categories 5 and 6) wines. For the example with 5 features, the respective numbers were 76.9% and 85.3%, which are identical to the outcome when using 8 features.

VI. Conclusions

With a dataset consisting of 180 Washington State Cabernet Sauvignon wine samples and 32 associated features, we have created a complete cost-based wine quality prediction tool. The model is optimal within a wide range of Weka implemented classifiers and feature ranking methods.

TABLE III
OPTIMAL COST-ACCURACY ACHIEVED AND THE NUMBER OF FEATURES SELECTED.

Total cost	Accuracy	# features
$ 24	60.55%	5
$ 77	66.66%	8
$105	66.66%	13
$148	70.00%	18
$198	71.11%	20
$456	73.33%	27
$756	74.44%	32

With respect to the number of features, our model is generated in linear time. The greedy approach provides a pseudo-optimal solution to an otherwise exponential time multi-criterial optimization problem.

Cost optimization with the fewest features and significant prediction accuracy in particular at the two end-members of wine quality is achieved with 5 features. In order of importance, the first and fourth are easily measurable colorimetric wine characteristics, the second and third are age and region of wine production, and the last is total SO_2 content, which is an anti-oxidant and anti-microbial added to wines.

These results show that a relatively simple multi-criterial optimization tool allows for an economical way to discriminate between high and low quality wines, without the need for an expert taster. Additional analyses and associated costs can provide further resolution in the prediction of the quality of mid-range wines.

REFERENCES

[1] D. Smith and R. Margolskee, "Making sense of taste," *Scientific American*, vol. 16, no. 3, pp. 84–92, 2006.

[2] A. Legin, A. Rudnitskaya, L. Lvova, Y. Vlasov, C. D. Natale, and A. D'Amico, "Evaluation of italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception," *Analytica Chimica Acta*, vol. 484, no. 1, pp. 33–44, 2003.

[3] N. H. Beltran, M. A. Duarte-Mermoud, 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," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 11, pp. 2421–2436, Nov 2008.

[4] D. Kruzlicova, J. Mocak, B. Balla, J. Petka, M. Farkova, and J. Havel, "Classification of slovak white wines using artificial neural networks and discriminant techniques," *Food Chemistry*, vol. 112, no. 4, pp. 1046–1052, 2009.

[5] A. Bednarova, R. Kranvogl, D. Voncina, T. Jug, and E. Beinrohr, "Characterization of Slovenian wines using multidimensional data analysis from simple enological descriptors," *Acta Hhim. Slov.*, vol. 60, pp. 274–286, 2013.

[6] A. Cichelli, F. Damiani, F. Murmura, M. S. Simonetti, M. Odoardi, and P. Damiani, "Classification of Montepulciano d'Abruzzo wines by linear discriminant analysis and artificial neural networks," *American Journal of Enology and Viticulture*, vol. 51, no. 2, pp. 108–114, 2000.

[7] C. Diaz, J. E. Conde, D. Estvez, S. J. P. Olivero, and J. P. P. Trujillo, "Application of multivariate analysis and artificial neural networks for the differentiation of red wines from the canary islands according to the island of origin," *Journal of Agricultural and Food Chemistry*, vol. 51, no. 15, pp. 4303–4307, 2003.

[8] A. Hosu, V.-M. Cristea, and C. Cimpoiu, "Analysis of total phenolic, flavonoids, anthocyanins and tannins content in romanian red wines: Prediction of antioxidant activities and classification of wines using artificial neural networks," *Food chemistry*, vol. 150, pp. 113–118, 2014.

[9] S. Perez-Magarino, M. Ortega-Heras, M. G.-S. Jose, and Z. Boger, "Comparative study of artificial neural network and multivariate methods to classify spanish DO rose wines," *Talanta*, vol. 62, no. 5, pp. 983–990, 2004.

[10] S. Gomez-Meire, C. Campos, E. Falque, F. Diaz, and F. Fdez-Riverola, "Assuring the authenticity of northwest spain white wine varieties using machine learning techniques," *Food Research International*, vol. 60, pp. 230–240, 2014.

[11] J. Aires-De-Sousa, "Verifying wine origin: A neural network approach," *American Journal of Enology and Viticulture*, vol. 47, no. 4, pp. 410–414, 1996.

[12] M. Gaeta, M. Marsella, S. Miranda, and S. Salerno, "Using neural networks for wine identification," *Intelligence and Systems, IEEE International Joint Symposia on*, vol. 0, p. 418, 1998.

[13] J. G. Ferrier and D. E. Block, "Neural-network-assisted optimization of wine blending based on sensory analysis," *American Journal of Enology and Viticulture*, vol. 52, no. 4, pp. 386–395, 2001.

[14] J. Ren and Z. Li, *A System of Wine Blending Based on Neural Network*. Boston, MA: Springer, 2006, pp. 604–609.

[15] H. Baykal and H. K. Yildirim, "Application of artificial neural networks (ANNs) in wine technology," *Critical Reviews in Food Science and Nutrition*, vol. 53, no. 5, pp. 415–421, 2013.

[16] A. Yamazaki, T. B. Ludermir, and M. C. P. de Souto, "Classification of vintages of wine by artificial nose using time delay neural networks," *Electronics Letters*, vol. 37, no. 24, pp. 1466–1467, 2001.

[17] R. Chandra, K. Chaudhary, and A. Kumar, "The combination and comparison of neural networks with decision trees for wine classification," School of Sciences and Technology, The University of Fiji, Lautoka, Fiji, Tech. Rep., 2007.

[18] P. Appalasamy, A. Mustapha, N. D. Rizal, F. Johari, and A. F. Mansor, "Classification-based data mining approach for quality control in wine production," *Journal of Applied Sciences*, vol. 12, pp. 598–601, 2012.

[19] A. Urtubia, J. R. Perez-Correa, A. Soto, and P. Pszczolkowski, "Using data mining techniques to predict industrial wine problem fermentations," *Food Control*, vol. 18, no. 12, pp. 1512–1517, 2007.

[20] F. Liu, L. Wang, and Y. He, "Application of least squares support vector machines for discrimination of red wine using visible and near infrared spectroscopy," in *Intelligent System and Knowledge Engineering (ISKE), 3rd International Conference on*, vol. 1, Nov 2008, pp. 1002–1006.

[21] H. Yu, H. Lin, H. Xu, Y. Ying, B. Li, and X. Pan, "Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy," *Journal of agricultural and food chemistry*, vol. 56, no. 1, pp. 307–313, 2008.

[22] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decis. Support Syst.*, vol. 47, no. 4, pp. 547–553, Nov. 2009.

[23] A. Nachev and M. Hogan, "Using data mining techniques to predict product quality from physicochemical data," *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, p. 1, 2013.

[24] H. Yu, X. Niu, H. Lin, Y. Ying, B. Li, and X. Pan, "A feasibility study on on-line determination of rice wine composition by Vis-NIR spectroscopy and least-squares support vector machines," *Food Chemistry*, vol. 113, no. 1, pp. 291–296, 2009.

[25] B. Zoecklein, *Wine Analysis and Production*, ser. Systematics Association Special Volume Series. Chapman & Hall, 1995.

[26] J. F. Harbertson and S. Spayd, "Measuring phenolics in the winery," *American Journal of Enology and Viticulture*, vol. 57, no. 3, pp. 280–288, 2006.

[27] J. Levengood and R. Boulton, *The Variation in the Color Due to Copigmentation in Young Cabernet Sauvignon Wines*. Washington, DC: American Chemical Society, 2004, ch. 5, pp. 35–52.

[28] J. C. De Man, M. Rogosa, and M. E. Sharpe, "A medium for the cultivation of lactobacilli," *Journal of Applied Bacteriology*, vol. 23, no. 1, pp. 130–135, 1960.

[29] J. Swings, M. Gillis, and K. Kersters, "Phenotypic identification of acetic acid bacteria," *Applied and Environmental Microbiology*, vol. 29, pp. 103–110, 1992.

[30] N. Rodrigues, G. Gonalves, S. Pereira-da Silva, M. Malfeito-Ferreira, and V. Loureiro, "Development and use of a new medium to detect yeasts of the genera dekkera/brettanomyces," *Journal of Applied Microbiology*, vol. 90, no. 4, pp. 588–599, 2001.

[31] A. Delaherche, O. Claisse, and A. Lonvaud-Funel, "Detection and quantification of brettanomyces bruxellensis and ropypediococcus damnosus

strains in wine by real-time polymerase chain reaction," *Journal of Applied Microbiology*, vol. 97, no. 5, pp. 910–915, 2004.

[32] E. T. Neeley, T. G. Phister, and D. A. Mills, "Differential real-time PCR assay for enumeration of lactic acid bacteria in wine," *Applied and environmental microbiology*, vol. 71, no. 12, pp. 8954–8957, 2005.

[33] R. S. Jackson, "11 - sensory perception and wine assessment," in *Wine Science (Second Edition)*, third edition ed., ser. Food Science and Technology, R. S. Jackson, Ed. San Diego: Academic Press, 2008, pp. 544 – 590.

[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[35] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. da Fontoura Costa, "A systematic comparison of supervised classifiers," *PLoS ONE*, vol. 9, no. 4, pp. 1–14, 04 2014.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[37] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on*, Nov 1995, pp. 388–391.

[38] P. Paclík, R. P. W. Duin, G. M. P. van Kempen, and R. Kohlus, "On feature selection with measurement cost and grouped features," pp. 461–469, 2002.

[39] V. Boln-Canedo, I. Porto-Daz, N. Snchez-Maroo, and A. Alonso-Betanzos, "A framework for cost-based feature selection," *Pattern Recognition*, vol. 47, no. 7, pp. 2481 – 2489, 2014.

[40] W. Shu and H. Shen, "Multi-criteria feature selection on cost-sensitive data with missing values," *Pattern Recognition*, vol. 51, pp. 268 – 280, 2016.