Assessing wine quality using a decision tree

Seunghan Lee, Juyoung Park and Kyungtae Kang*
Dept. of Computer Science & Engineering
Hanyang University, Republic of Korea
Email: {shlee10, parkjy, ktkang}@hanyang.ac.kr

***Abstract*—Even though wine-drinkers generally agree that wines may be ranked by quality, wine-tasting is famously subjective. There have been many attempts to construct a more methodical approach to the assessment of wines. We propose a method of assessing wine quality using a decision tree, and test it against the wine-quality dataset from the UC Irvine Machine Learning Repository. Results are 60% in agreement with traditional assessment techniques.**

I. Introduction

In response to the increasing popularity of wine and its adoption in new markets, the wine industry is improving its technologies for both wine making and sales [1]. Definitive statements about wine quality are notoriously difficult to make; but more objective metrics are key to expansion into a wider market, in which new wine drinkers seek reassurance on the merit of a wine.

The quality of a wine is generally assessed by sensory and physiochemical methods. The taste of a trained panelist currently gives more informed results, but this method of quality measurement is time-consuming and expensive. Physical and chemical tests are more repeatable, and are routinely used to characterize wine through measurements of density, sugar, tannin, alcohol and acidity levels. We propose a new approach, in which a decision tree [2], [3] is used to infer quality, as perceived by a wine drinker, from physiochemical characteristics. We test this approach on the Wine Quality Data Set from the UC Irvine Machine Learning Repository [4].

A few applications of data-mining techniques to wine quality assessment have already been proposed. Cortez *et al.* [1] attempted to predict taste preferences by applying a support vector machine, multiple regression, and a neural network to chemical analysis of wines. Shanmuganathan *et al.* [5] predicted the effects of season and climate on vine yields and wine quality. The Wineinformatics system due to Chen *et al.* [6] abstracts the flavor and characteristics of wine from natural-language reviews, by applying hierarchical clustering and association rules.

II. Method

A. Wine Data

The Wine Quality Data Set from the UCI Repository [4] contains data for 1599 red wines and 4898 white wines from Portugal, each of which is associated with 11 attributes. Table I summarizes the scope of this physiochemical data. Every wine in the data set has also been evaluated by a minimum of three experts in a blind tasting, and graded from 0 (very bad) to 10 (excellent). Fig. 1 shows the importance of each item of physiochemical data in the UCI Repository [1].

TABLE I. Means and Ranges of the Physicochemical Data in the UCI Wine Quality Data Set

Attribute	Red wine Mean (Range)	White wine Mean (Range)
Fixed acidity	8.3 (4.6 - 15.9)	6.9 (3.8 - 14.2)
Volatile acidity	0.5 (0.1 - 1.6)	0.3 (0.1 - 1.1)
Citric acid	0.3 (0.0 - 1.0)	0.3 (0.0 - 1.7)
Residual sugar	2.5 (0.9 - 15.5)	6.4 (0.6 - 65.8)
Chlorides	0.08 (0.01 - 0.61)	0.05 (0.01 - 0.35)
Free sulfur dioxide	14 (1 - 72)	35 (2 - 289)
Total sulfur dioxide	46 (6 - 289)	138 (9 - 440)
Density	0.996 (0.990 - 1.004)	0.994 (0.987 - 1.039)
pH	3.3 (2.7 - 4.0)	3.1 (2.7 - 3.8)
Sulphates	0.7 (0.3 - 2.0)	0.5 (0.2 - 1.1)
Alcohol	10.4 (8.4 - 14.9)	10.4 (8.0 - 14.2)

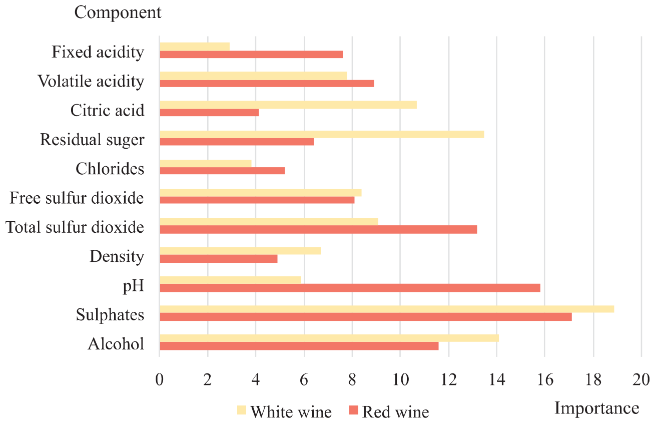


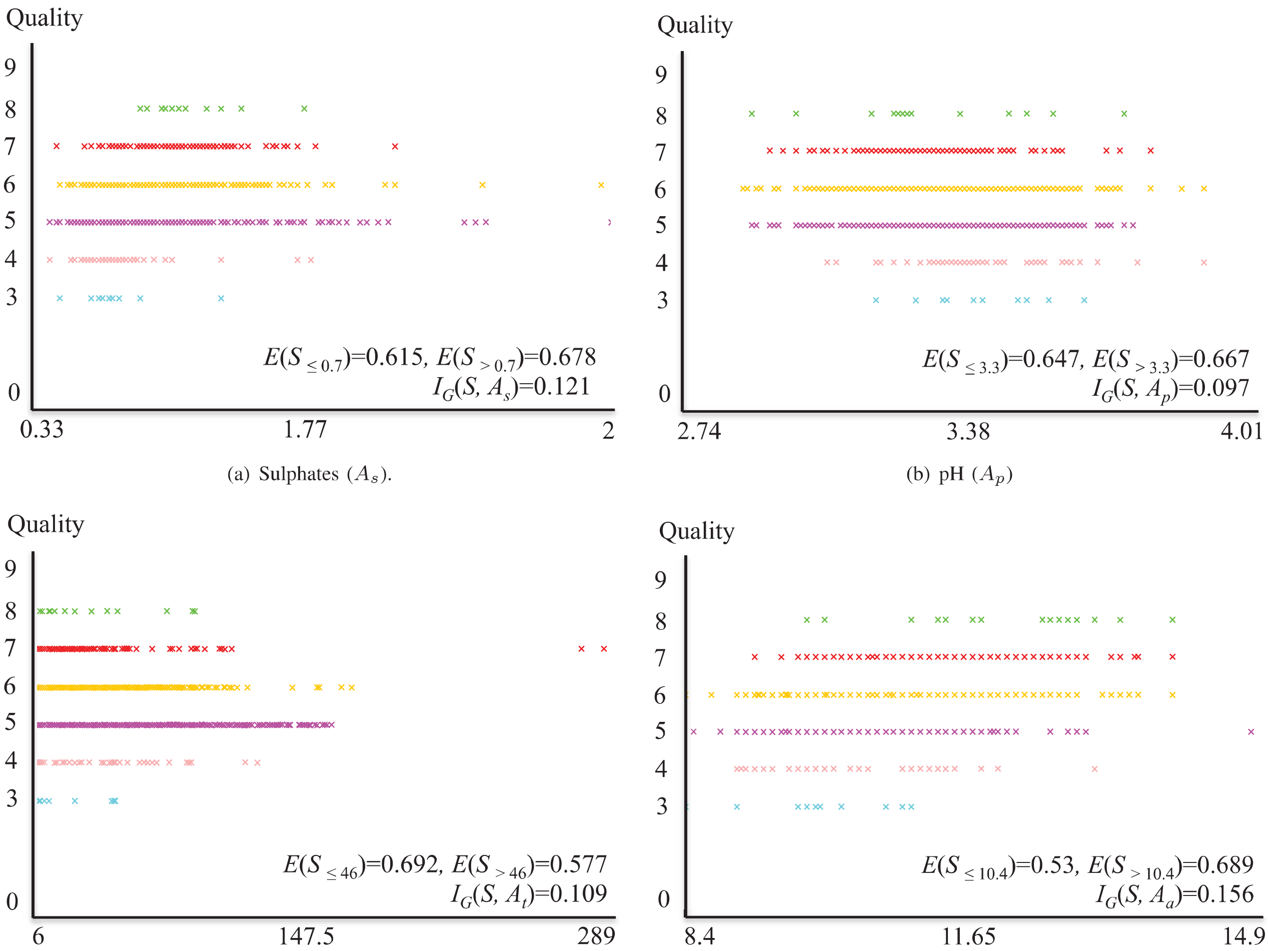
Fig. 1. Importance of physiochemical indicators.

B. Using the Decision Tree to Predict Wine Quality

Our predictive model applies the C4.5 [7] algorithm to build a decision tree [2], based on physiochemical characteristics of wine, and uses it to predict taste preferences. C4.5 constructs a decision tree by recursive subdivision, selecting the most influential attribute from the training instances at each node, and splitting that set of instances into subsets with high

*Corresponding author

978-1-4799-1920-8/15/$31.00 ©2015 IEEE



(a) Sulphates (A_s).

(b) pH (A_p)

(c) Total sulfur dioxide (A_t)

(d) Alcohol (A_a)

Fig. 2. Effect of key physiochemical indicators on the quality of red wines.

and low values of that attribute; each of these subsets then becomes a child node. After subdivision is complete, all the instances are distributed across the leaf nodes.

To select the most influential attribute, C4.5 uses information gain, considered as a reduction in entropy ($E(S)$) [8], [9], which is expressed as $\sum_{i=1}^{n}(-p_i \log_2 p_i)$, where S denotes a set of training instances with their 11 attributes, i denotes the degree of preference, and p_i denotes the proportion of S with degree i. If all the instances belong to the same class, the value of $E(S) = 0$, if they all belong to different classes, then $E(S) = 1$. We can go on to express the effectiveness of an attribute as an information gain I_G, as follows:

$$I_G(S, A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} E(S_v), \qquad (1)$$

where $V(A)$ denotes the set of all possible values of the attribute A, and S_v is the subset of S in which the value of the attribute is v. The most influential attribute at a node is that with the highest information gain.

To illustrate this approach, we compared the entropy and information gain of four key physiochemical indicators of the quality of red wines, with the results shown in Fig. 2. First, the entropy $E(S)$ of all instances was computed, and found to be 0.754. Then the information gain of each attribute was computed using Eq. (2), while setting v to the mean value of that attribute in this group of instances. We see that perceived quality is closely linked to these measures with alcohol level yielding the highest information gain, by a short head. This suggests that alcohol level is the most significant determinant of the perceived quality of red wine.

III. Results

Table II presents the results of a five-fold cross-validation [10] of our predictions of taste preferences for wines. The measured quality was from 3-8 for red wine and 4-8 for white wine. Our predictions for red wines have a precision of 61.1%, a recall of 60.7%, an f-measure of 60.3%, and an accuracy of 60.7%, on average. For white wines, the corresponding averages were: precision 58.2%, recall 58.7%, F-measure 58.4%, and accuracy 58.7%.

We then compared our model with results from Weka's [11] implementation of three machine-learning algorithms (LibSVM, BayesNet, MultiPerceptron), applied to the same data,

TABLE II. ACCURACY OF PREDICTED TASTE PREFERENCES

Quality	White wine			Red wine		
	Precision (%)	Recall (%)	F-Me (%)	Precision (%)	Recall (%)	F-Me (%)
3	7.7	10.0	8.70	–	–	–
4	24.4	20.8	22.4	27.3	23.9	25.5
5	48.2	72.2	70.2	60.0	61.9	60.9
6	57.9	57.7	57.8	63.9	64.6	64.3
7	55.7	48.7	52.0	52.8	51.9	52.4
8	10.0	60.7	60.3	36.7	31.4	33.8
Avg	61.1	60.7	60.3	58.2	58.7	58.5
Ac (%)	60.7			58.7		

Avg: Average, Ac: Accuracy, F-Me: F-Measure

TABLE III. COMPARATIVE PERFORMANCE OF MACHINE-LEARNING ALGORITHMS

Prediction model	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
LibSVM	55.70	58.00	55.70	58.03
BayesNet	56.40	58.30	56.90	58.28
MultiPerceptron	57.70	59.70	58.30	59.66
Proposed	60.10	60.70	60.30	60.66

in terms of precision, recall, F-measure, and accuracy, again using five-fold cross-validation. Table III clearly shows that our model outperforms the others.

IV. CONCLUSION

We have proposed a new way of predicting taste preferences for wines using a decision tree, and evaluated it using the Wine Quality Data Set from the UCI Machine Learning Repository. Results suggest that this model could offer untutored consumers a better chance of selecting a high-quality wine.

ACKNOWLEDGMENT

This work was partly supported by a grant from the Institute for Information & Communications Technology Promotion (IITP), funded by the Korean government (MSIP) (No. B0101-15-0557, Resilient Cyber-Physical Systems Research), and partly supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the MSIP (NRF-2013R1A1A1059188).

REFERENCES

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

[2] A. Abdelhalim and I. Traore, “A new method for learning decision trees from rules,” *IEEE International Conference on Machine Learning and Applications*, pp. 693–698, Dec. 2009.

[3] M. A. Hussain, M. K. Rao, and A. M. Mahmood, “An optimized approach to generate simplified decision trees,” *IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–5, Dec. 2013.

[4] A. Asuncion and D. Newman, UC Irvine Machine Learning Repository, [Online] Available: http://archive.ics.uci.edu/ml/index.html.

[5] S. Shanmuganathan, P. Sallis, and A. Narayanan, “Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality,” *IEEE International Conference on Computational Intelligence, Communication Systems and Networks*, pp. 82–89, July 2010.

[6] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, “Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel,” . *IEEE International Conference on Data Mining Workshop*, pp. 142–149, Dec. 2014.

[7] J. R. Quinlan, “Improved use of continuous attributes in C4.5,” *Journal of Artificial Intelligence Research 4*, vol. 4, no. 1, pp. 77–90, Jan. 1996.

[8] I. M. Mitchell, *Machine Learning*, McGraw-Hill International Editions, 1997.

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann; Second Edition, 2005.

[10] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” In *Proc. International Joint Conference on Artificial Intelligence*, pp. 1137–1143, Aug. 1995.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 1997.