Journal of Applied Sciences 12 (6): 598-601, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.598.601

Classification-based Data Mining Approach for Quality Control in Wine Production

P. Appalasamy, A. Mustapha, N.D. Rizal, F. Johari and A.F. Mansor
Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: Modeling the complex human taste is an important focus in wine industries. The main purpose of this study was to predict wine quality based on physicochemical data. This study was also conducted to identify outlier or anomaly in sample wine set in order to detect adulteration of wine. In this project, two large separate datasets are used, which contains 1, 599 instances for red wine and 4, 989 instances for white wine with 11 attributes of physicochemical data such as alcohol, PH and sulfates. Two classification algorithms, Decision tree and Naïve Bayes are applied on the dataset and the performance of these two algorithms is compared. Results showed that Decision tree (ID3) outperformed Naïve Bayesian techniques particularly in red wine, which is the most common type. The study also showed that two attributes, alcohol and volatile-acidity contribute highly to wine quality. White wine is also more sensitive to changes in physicochemistry as opposed to red wine, hence higher level of handling care is necessary. This research concludes that classification approach will give rooms for corrective measure to be taken in effort to increase the quality of wine during production.

Key words: Classification, wine quality, Naïve Bayes, decision tree

INTRODUCTION

Wine industry is currently growing well in the market since the last decade. However, the quality factor in wine has become the main issue in wine making and selling. To meet the increasing demand, assessing the quality of wine is necessary for the wine industry to prevent tampering of wine quality as well as maintaining it. To remain competitive, wine industry is investing in new technologies like data mining for analyzing taste and other properties in wine. Data mining techniques provide more than summary, but valuable information such as patterns and relationships between wine properties and human taste, all of which can be used to improve decision making and optimize chances of success in both marketing and selling (Cortez *et al*., 2009).

Two key elements in wine industry are wine certification and quality assessment, which are usually conducted via physicochemical and sensory tests (Ebeler, 1999). Physicochemical tests are lab-based and are used to characterize physicochemical properties in wine such as its density, alcohol or pH values. Physicochemical properties in wine have been widely studied in wine production fermentation (Reddy *et al*., 2006), apricot and raisin wine (Bapat *et al*., 2010), brewed red wine (Samappito and Butkhup, 2008), Thai rice wine (Sirisantimethakom *et al*., 2008), palm wine (Ogbo *et al*., 2009) as well as in Nigerian local beverages (Adeleke and Abiodun, 2010).

Meanwhile, sensory tests such as taste preference are performed by human experts. Taste is a particular property that indicates quality in wine, the success of wine industry will be greatly determined by consumer satisfaction in taste requirements. Physicochemical data are also found useful in predicting human wine taste preference (Cortez *et al*., 2009) and classifying wine based on aroma chromatograms (Beltran *et al*., 2008). In predicting human wine taste preference, Cortez and his colleagues uses three regression techniques, which are the Support Vector Machine (SVM), multiple regression, and neural networks from machine learning. Each sample in this dataset was first evaluated by a minimum of three sensory assessors (using blind tastes) who graded the wine according to a scale that ranges from 0, which is very bad to 10, which is excellent.

Another work involves wine classification based on the physicochemical information contained in wine aroma chromatograms as measured with a Fast GC Analyzer (Beltran *et al*., 2008). This study compares the performance of three classification methods, which are

Corresponding Author: Aida Mustapha, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia Tel: +603-8947 1714

Linear Discriminate Analysis (LDA), Radial Basis Function Neural Networks (RBFNN), and Support Vector Machines (SVM) in a two-staged architecture.

Following Cortez (Cortez *et al*., 2009), the main objectives of this study were to predict wine quality using the same physicochemical data but to compare using two different algorithms, which are the decision tree-based ID3 and Naïve Bayesian. Next is to find out pattern in attributes that affect the quality of wine.

MATERIALS AND METHODS

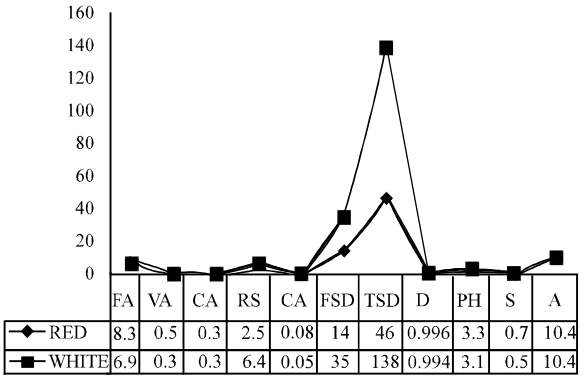
In achieving the objectives of this study, dataset preprocessing and classification experiments are conducted using WEKA (Hall *et al*., 2009) on a Windows PC with Duo-Core 1.83 GHz CPU and 4 GB RAM. The dataset is a wine quality dataset that is publicly available for research purposes from http://www3.dsi.uminho.pt/pcortez/wine/ (Cortez *et al*., 2009). It consists of two separate datasets, red wine and white wine. Both dataset contains 1,599 instances with 11 attributes for red wine and 4, 989 instances and the same 11 attributes for white wine. Each instance is classified into quality attribute that ranges between 0 (very bad) and 10 (excellent).

The attributes include Fixed Acidity (FA) in g (tartaric acid) dm^{-3}, Volatile Acidity (VA) in g (acetic acid) dm^{-3}, Citric Acid (CA) in g dm^{-3}, Residual Sugar (RS) in g dm^{-3}, Chlorides (C) in g (sodium chloride) dm^{-3}, Free Sulfur Dioxide (FSD) in mg dm^{-3}, Total Sulfur Dioxide (TSD) in mg dm^{-3}, Density (D) in g cm^{-3}, pH, Sulphates (S) in g (potassium sulphate) dm^{-3}, and Alcohol (A) in volume percentage. Figure 1 shows the statistical mean for all physicochemical attributes in both wine types.

From Fig. 1, the maximum mean value is 138 mg dm^{-3} for Total Sulfur Dioxide (TSD) in white wine as compared to only 46 mg dm^{-3} of TSD in red wine.

Preprocessing: Data preprocessing is an important step before applying any data mining techniques on the dataset. For this study, only discretization and attribution selection are performed as the other processes are not relevant for the dataset. This dataset has already been cleaned therefore there is no more need to apply data cleaning. During discretization, all continuous value in the dataset is converted into discrete data. For this study, both datasets are automatically discretized from a range of numeric attributes into nominal by using first-last method in WEKA. After the discretization process, the number of attributes becomes smaller than the attributes in the original dataset.

The second data preprocessing is attribute selection. In this process, relevant data are selected in order to reduce data dimensionality that affects computational cost in data mining. The selection of attributes is performed automatically by WEKA using Info Gain Attribute Eval method. The method evaluates the worth of an attribute by measuring the information gain with respect to the class. For this dataset, it is found out that some attribute are highly correlated with each other and attribute selection is necessary to reduce complexity during data mining process. In preprocessing stage, we apply attribute selection algorithm which resulted in attribute ranking for each dataset as shown in Table 1.



	FA	VA	CA	RS	CA	FSD	TSD	D	PH	S	A
RED	8.3	0.5	0.3	2.5	0.08	14	46	0.996	3.3	0.7	10.4
WHITE	6.9	0.3	0.3	6.4	0.05	35	138	0.994	3.1	0.5	10.4

Fig. 1: Statistical mean for each physicochemical attributes in both wine types. The attributes include, FA: Fixed Acidity in g (tartaric acid) dm^{-3}, VA: Volatile Acidity in g (acetic acid) dm^{-3}, CA: Citric Acid in g dm^{-3}, RS: Residual Sugar in g dm^{-3}, C: Chlorides in g (sodium chloride) dm^{-3}, FSD: Free Sulfur Dioxide in mg dm^{-3}, TSD: Total Sulfur Dioxide in mg dm^{-3}, D: Density in g cm^{-3}, pH, S: Sulphates in g (potassium sulphate) dm^{-3}, and A: Alcohol in volume percentage

Table 1: Ranking attributes using attribute selection algorithm

Red wine dataset	White wine dataset
Volatile Acidity (2)	Volatile Acidity (2)
Total Sulfur Dioxide (7)	Citric Acid (3)
Sulphate (10)	Chlorides (5)
Alcohol (11)	Free Sulfur Dioxide (6)
	Density (8)
	Alcohol (11)

Classification: The objective of this classification experiment was to investigate physicochemistry properties in wine that influence the taste, hence the quality of a wine. Two classification algorithms applied on the dataset are ID3 from the Decision Tree family and Naïve Bayesian from Bayesian theory. Both datasets are separated into training and testing set by using 10-fold cross-validation method. The training data is randomly portioned into 10 sets of equal size and the algorithms are executed 10 times.

Decision tree-: The first algorithm chosen is decision tree using ID3 method. The ID3 method gives the highest accuracy among other decision tree's methods such as random tree, FT (Functional Tree), NBTree, and simple cart in WEKA. This method builds the tree from top down, with no backtracking. Furthermore, by using this method, the tree is built faster and the leaves are shorter when compared to other methods.

Naïve Bayes: The second algorithm is Naïve Bayes algorithm that is able to find two highly correlated attributes that is similar to attribute selection purpose. By using Naïve Bayes algorithm, the model is built faster and it is highly scalable as compared to other Bayes' algorithms in WEKA.

RESULTS AND DISCUSSION

The classification experiment is measured by accuracy percentage of classifying the instances correctly into its class according to quality attributes ranges between 0 (very bad) and 10 (excellent). From the experiments, we found that classification for red wine quality using ID3 algorithm achieved 60.0% accuracy while Naïve Bayesian classifier achieved about 58.8% accuracy. For the white wine, ID3 algorithm yielded 52.3% accuracy while Naïve Bayesian classifier yielded 50.5% accuracy. Efficiency of ID3 algorithm in terms of processing time is logged ranging from 30-60 sec, while Naïve Bayesian classifier is found to be 1.5 to 2.5 min.

The experimental results are summarized in Table 2 and are compared against the accuracy percentage using support vector machine by Cortez *et al.* (2009). Detailed results that produce the accuracy percentage in Table 2 for both ID3 and Naïve Bayesian classifier for-red wine and white wine are presented by confusion matrices as shown in Table 3 and 4, respectively.

Results from the experiments lead us to conclude that ID3 performs better in classification task as compared against the Naïve Bayesian classifier. The processing time for ID3 algorithm is also observed to be more efficient and less time consuming despite the large size of wine properties dataset. We found that human taste is too complex to model correctly using simple classification techniques such as ID3 and Naïve Bayesian as compared to SVM by Cortez *et al.* (2009).

Our main observation from the research we conducted is that ID3 and Naïve Bayesian models are not much useful in classifying finished wine product due to considerably low accuracy, whereby we misclassify the wine. Nonetheless, our intention is to provide a lightweight comparative model in addition to the SVM model by Cortez *et al.* (2009). Classification models may be used as part of decision support system in different stages of wine production, hence giving the opportunity for manufacturer to make corrective and additive measure that will result in higher quality wine being produced.

Table 2: Wine modeling results

Correctly classified instances	ID3 (%)	Naïve Bayesian (%)	SVM (%)
Red Wine	60.0	58.8	62.4
White Wine	52.3	50.5	64.6

Table 3: Confusion matrices for red wine and white wine using ID3

	Red wine					White wine						
Class	4	5	6	7	8	3	4	5	6	7	8	9
3	1	7	2	0	0	2	3	7	6	0	0	0
4	4	29	19	0	0	4	39	63	46	4	0	0
5	8	498	169	3	0	9	43	767	549	55	2	0
6	5	200	405	25	1	4	23	481	1400	247	4	1
7	0	8	137	53	0	0	3	58	463	345	4	0
8	0	0	10	8	0	0	0	8	83	70	11	0
9	0	0	0	0	0	0	0	0	3	1	1	0

Table 4: Confusion matrices for red wine and white wine using Naïve Bayesian

	Red wine					White wine						
Class	4	5	6	7	8	3	4	5	6	7	8	9
3	3	6	1	0	0	1	2	7	8	2	0	0
4	6	30	17	0	0	1	22	86	42	11	1	0
5	6	481	185	9	0	2	16	935	430	68	0	0
6	5	193	373	67	0	0	8	634	1058	488	10	0
7	0	12	106	81	0	0	0	69	356	450	5	0
8	0	0	8	10	0	0	1	10	59	99	6	0
9	0	0	0	0	0	0	0	0	1	4	0	0

From the resulting classification accuracy, we found that accuracy rate for the white wine is influenced by a higher number of physicochemistry attribute, which are alcohol, density, free sulfur dioxide, chlorides, citric acid, and volatile acidity. Meanwhile, red wine quality is highly correlated to only four attributes, which are alcohol, sulphates, total sulfur dioxide, and volatile acidity. This shows white wine quality is affected by physicochemistry attributes that does not affect the red wine in general. Therefore, we suggest that white wine manufacturer should conduct wider range of test particularly towards density and chloride content since white wine quality is affected by such substances.

CONCLUSION

Since, white wine is more sensitive to changes in physicochemistry properties as compared to red wine, we suggest a higher level of separation between white wine and red wine production line with particularly further customization to the white wine production. Attribute selection algorithm we conducted also ranked alcohol as the highest in both datasets, hence the alcohol level is the

main attribute that determines the quality in both red and white wine. Our suggestion is that wine manufacturer to focus in maintaining a suitable alcohol content, may be by longer fermentation period or higher yield fermenting yeast.

REFERENCES

Adeleke, R.O. and O.A. Abiodun, 2010. Physico-chemical properties of commercial local beverages in Osun State, Nigeria. Pak. J. Nutr., 9: 853-855.

Bapat, R.K., S.B. Jadhav and J.S. Ghosh, 2010. Fermentation and characterization of apricot and raisin Wine by *Saccharomyces cerevisiae* NCIM 3282. Res. J. Microbiol., 5: 1093-1099.

Beltran, N.H., M.A. Duarte-Mermound, V.A.S. Vicencio, S.A. Salah and M.A. Bustos, 2008. Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer. Instrum. Measurement. IEEE Trans., 57: 2421-2436.

Cortez, P., A. Cerdeira, A. Fernando, M. Telmo and R. Jose, 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Syst., 47: 547-553.

Ebeler, S., 1999. Flavor Chemistry-Thirty Years of Progress: Linking Flavor Chemistry to Sensory Analysis of Wine. Kluwer Academic Publishers, Norwell, MA, USA., pp: 409-422.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, 2009. The WEKA data mining software: An update. SIGKDD Explorations Newslett., 11: 10-18.

Ogbo, F.C., J.A. Onuegbu and O.K. Achi, 2009. Improvement of protein content of garri by inoculation of cassava mash with biomass from palm wine. Am. J. Food Technol., 4: 60-65.

Reddy, L.V.A., Y.H.K. Reddy and O.V.S. Reddy, 2006. Wine production by guava piece immobilized yeast from Indian cultivar grapes and its volatile composition. Biotechnology, 5: 449-454.

Samappito, S. and L. Butkhup, 2008. An analysis on flavonoids, phenolics and organic acids contents in brewed red wines of both non-skin contact and skin contact fermentation techniques of Mao luang ripe fruits (*Antidesma bunius*) harvested from Phupan valley in Northeast Thailand. Pak. J. Biol. Sci., 11: 1654-1661.

Sirisantimethakom, L., L. Lakkana, D. Paiboon and L. Pattana, 2008. Olatile compounds of a traditional Thai rice wine. Biotechnology, 7: 505-513.