**Classification based data mining for wine quality evaluation and predication**

CSC7442 group project report; Professor, Dr. Jianhua Chen

Chen Fang, Yun Ding, Chenguang Zhang, Jie Chen, Xinan Zhang and Haikuo Yu

**Abstract** – Wine quality evaluation and prediction are sometimes expensive and time consuming, which requires assessment by human experts’ knowledge and experience. This study explores a data mining based approach for wine quality evaluation and prediction and an interactive user interface for wine quality assessment and prediction was provided. The physicochemical characteristics of red and white wine samples were used to build the model and assess the result. 7 data mining techniques were investigated in the wine quality and predication domain, which are Naive Bayes, Logistic Regression, Linear Discriminant Analysis (LDA), Linear Support Vector Clustering (LSVC), Decision Tree, Support Vector Clustering (SVC) and Boosting. The results show that (we will update this part later according to the results.)

**1 Introduction**

Wine produce and consumption are an important industrial area, and the current technologies have improved the whole wine produce efficiency and process, while wine quality is still heavily dependent on the experts’ taste and evaluation. With the increase of demand for wine, an effective wine quality evaluation and prediction is essential for wine industrial, which may boost wine produce and consumption ([Cortez et al. 2009](#_ENREF_3), [Tian and Pang 2010](#_ENREF_8), [Appalasamy et al. 2012](#_ENREF_1)). Wine certification and quality assessment, as two key elements in wine industry, are usually conducted via physicochemical and sensory test ([Ebeler 1999](#_ENREF_5)). The physicochemical test is lab-based evaluation ([Appalasamy et al. 2012](#_ENREF_1)) and it investigates the properties of wine, including wine density, alcohol or PH values. Meanwhile, sensory test is human experts-based evaluation, such as taster preference. With the rapid development of data mining technology, business intelligence becomes more and more important for companies to upgrade the quality of product ([Tian and Pang 2010](#_ENREF_8)).

Data mining has been widely applied in quality evaluation, and the overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Several data mining technologies were investigated for wine quality evaluation and prediction. Cortez et al. ([Cortez et al. 2009](#_ENREF_3)) investigated the physicochemical data in prediction human wine taste preference. Beltran et al. ([Beltran et al. 2008](#_ENREF_2)) classified wine quality based on aroma chromatograms with Fast GC Analyzer, and compared the performance of different classification methods, which are Linear Discriminate Analysis (LDA), Radial Basis Function Neural Networks (RBFNN) and Support Vector Machine ([Beltran et al. 2008](#_ENREF_2), [Appalasamy et al. 2012](#_ENREF_1)). Several other data mining technologies ([Nachev and Hogan](#_ENREF_6) , [Beltran et al. 2008](#_ENREF_2), [Cortez et al. 2009](#_ENREF_4), [Ribeiro et al. 2009](#_ENREF_7), [Tian and Pang 2010](#_ENREF_8), [Appalasamy et al. 2012](#_ENREF_1)) were applied in predicting human wine taste preference, such as Support Vector Machine, multiple regression and neural networks. Appalasamy et al. ([Appalasamy et al. 2012](#_ENREF_1)) applied Decision Tree and Naïve Bayes classification algorithms for wine quality control and discussed the pattern in attributes that affect the quality of wine. Their result ([Appalasamy et al. 2012](#_ENREF_1)) shows that Decision Tree (ID3) outperformed Naïve Bayes techniques in red wine. Tian and Pang 2010 ([Tian and Pang 2010](#_ENREF_8)) investigated three data mining algorithm, linear/multiple regression, decision tree, neural network, and discussed the advantages of each algorithm. Owing to the scale of different parameters of measured wine quality, Urtubia et al. ([Urtubia et al. 2007](#_ENREF_9)) used a two-stage classification procedure to predict wine problem fermentation, which includes PCA to reduce system dimension and K-Means to group fermentations into clusters of similar behavior.

Fewer data analysis is used in traditional wine industry ([Tian and Pang 2010](#_ENREF_8)), while data analysis including data mining technologies could process large and complex experimental data for accurate wine quality prediction([Tian and Pang 2010](#_ENREF_8)). The domain of this research was to explore a potential application of data mining technologies in wine evaluation and prediction. 7 data mining techniques were implemented to test the same wine data set and the results were compared with the corresponding classification accuracy, and they are Naive Bayes, Logistic Regression, Linear Discriminant Analysis (LDA), Linear Support Vector Clustering (LSVC), Decision Tree, Support Vector Clustering (SVC) and Boosting.

The experimented results shows that (we will update this part late according to the results)

**2 Data set**

The experiment data set was downloaded from Machine Learning Repository at University of California, Irvine ([Cortez et al. 2009](#_ENREF_3)). The data sets contain two parts, red and white vinho verde wine samples, from the north of Portugal. The attribute of data set is based on physicochemical tests and only physicochemical (inputs) and sensory (the output) variables are available for research usage, for example, no data about grape types, wine brand, wine selling price were provided in the data source. The data set is suitable for classification and regression. Although the wine classes are order, they are not balanced. Whether the input variable are relevant has not been confirmed in the metadata. There are total 12 variables ([Cortez et al. 2009](#_ENREF_3)) with 11 input variables and 1 output variable. The physicochemical tests based input variables include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and the sensory data based output variable is quality , which is a nominal score between 0 and 10.

**3 Materials and Methods**

1. **Machine learning models in this study**

We have trained and validated several machine learning classifying models in this study, including Gaussian Naïve Bayesian, Linear Discriminant Analysis, Logistic Regression, Linear Support Vector Machine, Decision Tree, and Support Vector Regression. The models mentioned above are from the open source Python machine-learning package, *Scikit-learn*[F. Pedrogesa]. In addition, we developed our in-house decision tree algorithm, which yields a relatively high accuracy compared with the well-optimized open source software.

1. **Hyperparameter optimization**

In order to obtain good generalization[J. Bergstra], hyperparamter optimization was used in this study. In contrast with the actual learning problems, which are often tries to optimize a loss function on the training set alone, hyperparamter optimization is often chosen to avoid the problem of overfitting. In this regard, we adopted accuracy on the validation dataset as our goal to guide the optimization.

1. **Ensemble method using a majority voting scheme**

The classification accuracy can be potentially improved by aggregating the predictions of multiple classifiers; this method is known as the ensemble method. The idea is to construct a set of base classifiers from training data and performs classification by taking a vote on the target value predicted by each base classifier[TBD]. Since we have already built and optimized several classifiers at hand, a simple majority voting scheme was implemented based on these classifiers in hope to harness the power of ensemble methods.

**4 Results and Discussion**

**5 Conclusions**

Effective wine quality evaluation, prediction and management are necessary to the rapid development of wine industry. As traditionally wine evaluation is labor intensive, the new data mining technologies may improve the whole wine evaluation process and then promote the wine industry. Several classification data mining technologies were adopted for wine quality evaluation and the each of them shows specific performance. A user friendly interface is also provided for users to evaluate the wine quality, which could be further used as reference or feedback for the wine industry to guide the wine quality.

The results show that (we will update this part according to result)

**References**

Appalasamy, P., et al. (2012). "Classiﬁcation-based Data Mining Approach for Quality Control in Wine Production." Journal of Applied Sciences **12**(6): 598-601.

Beltran, N. H., et al. (2008). "Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer." Instrumentation and Measurement, IEEE Transactions on **57**(11): 2421-2436.

Cortez, P., et al. (2009). "Modeling wine preferences by data mining from physicochemical properties." Decision Support Systems **47**(4): 547-553.

Cortez, P., et al. (2009). Using Data Mining for Wine Quality Assessment. Discovery Science. J. Gama, V. Costa, A. Jorge and P. Brazdil, Springer Berlin Heidelberg. **5808:** 66-79.

Ebeler, S. E. (1999). Linking flavor chemistry to sensory analysis of wine. Flavor Chemistry, Springer**:** 409-421.

Nachev, A. and M. Hogan "Using Data Mining Techniques to Predict Product Quality from Physicochemical Data."

Ribeiro, J., et al. (2009). Wine vinification prediction using data mining tools. Conference Proceedings, Computing and Computational Intelligence, Tbilisi, Republic of Georgia.

Tian, H. and Q. Pang (2010). Data mining application for upgrading quality of wine production. The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding.

Urtubia, A., et al. (2007). "Using data mining techniques to predict industrial wine problem fermentations." Food Control **18**(12): 1512-1517.