Jason Chan

Christopher Daffron (Team Lead)

Josh Willis

Pattern Recognition: Final Project

November 21st, 2014

Proposal for Classifying Wine Quality Data

Recently, the wine industry has been growing at a high rate. To support this high growth rate, the industry has invested in various technologies and techniques to create and sell wine [1]. Wine data collection and analysis can be used to prevent illegal adulteration of wines while ensuring quality in the market. Wine analysis is generally conducted on the results of a series of physicochemical and sensory tests [2].

Red and white vinho verde wine from Portugal was tested for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol [2]. The results of these tests describe wine quality based on chemical measurements. In addition to chemical measurements of wine, it would also be beneficial to measure the quality of wine by taste. Because taste is highly subjective, expert wine taste testers were used to score the quality of wine. This wine quality data set can be found from the UCI Machine Learning Repository [3]. Classifiers can predict the taste quality of the wine using the chemical measurements as predictors.

Neural networks (multilayer perceptrons) and support vector machines have been used to classify the quality of wine by the data collector, Cortez [2]. 5-fold cross validation was used to separate the data into training and testing sets. With a tolerance of 0.5, support vector machines

had an accuracy of 62.4% for red wine and 64.6% for white wine and neural networks had an accuracy of 59.1% for red wine and 52.6% for white wine [2]. Classification using neural networks and support vector machines is difficult because several parameters, such as the number of hidden nodes in a neural network or the kernel parameter in a support vector machine, must be correctly set to get good results. However, from the analyses we researched, support vector machines have been found to have the highest classification accuracy for the wine quality data set.

In contrast to classifications performed by Cortez, Appalasamy and company have classified wine quality using an ID3 decision-tree classifier and a Naive Bayes classifier. To use the ID3 and Naive Bayes classifiers, Appalasamy and company used 10-fold cross validation to break the data into training and testing sets. The ID3 classifier was found to have 60% accuracy for red wine and 52.3% accuracy for white wine [4]. The Naive Bayes classifier was found to have 58.8% accuracy for red wine and 50.5% for white wine.[4]. Appalasamy and company have "found that human taste is too complex to model correctly using simple classification techniques such as ID3 and Naive Bayesian" [4].

Since the data has been collected by Cortez and provided by the UCI Machine Learning Repository, we do not have to collect data on wine quality. First, we will decide how many of the features are significant and remove any insignificant features using principal component analysis (PCA) and Fisher's linear discriminant (FLD). Similar to Cortez and Appalasamy, we will use k-fold cross validation to split the data set into testing and training sets. Our classifiers will be run against these testing and training sets. We will use the classification techniques learned in this course, which includes: maximum posterior probability (MPP), k-nearest neighbor (kNN), backpropagation neural networks, decision trees, support vector machines (SVM), k-means, k-

means with winner-takes-all (WTA), and k-means with Kohonen self-organizing maps (SOM). Some of these classification techniques overlap with techniques already used on the wine quality data. In those cases, we will compare our results to the previously documented results. Finally, we will apply three of the four classifier fusion methods taught in this course, majority voting, Naive Bayes combination (NB), behavior-knowledge space (BKS), interval-based integration, to combine the results of the classification methods used.

References

[1] L. Jing et al., "The graphical feature extraction of star plot for wine quality classification," *First International Conference on Pervasive Computing, Signal Processing, and Applications,* pp. 771-774, 2010.

[2] P. Cortez et al., "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems,* no. 47, pp. 547-533, 2009.

[3] [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Wine+Quality. [Accessed 12 November 2014].

[4] P. Appalasamy et al., "Classification-based Data Mining Approach for Quality Control in Wine Production," *Journal of Applied Sciences,* vol. 12, no. 6, pp. 598-601, 2012.