CITS3401 Data Exploration and Mining
Project 2

Wine Classification

Mitchell Pomery
21130887

June 1, 2014

Introduction

The project specified that we are to develop several classifiers for wines using different classification methods to compare how machine learning performs compared to experts when rating different wines. The initial data is split into two groups, red wine and white wine, available from the UCI Machine Learning Repository[1]. Data analysis was done using Weka[2], data mining software created by Machine Learning Group at the University of Waikato, New Zealand. Specifically, we are using the classification tools in Weka Explorer for analysis.

Data Preprocessing

The initial data provided was in two files, `winequality-red.csv` and `winequality-white.csv`, that where converted to Weka's ARFF file format using an online conversion tool[3]. This tool was used to output two datasets, dataset 1 (`ds1-red.arff` and `ds1-white.arff`) and dataset 2 (`ds2-red.arff` and `ds2-white.arff`). Dataset 1 contains all the information that was in the original data, and is used to create the calssifier. The feilds in this dataset are numeric, apart from the quality which is nominal, making it is possible to group wines that recieve the same rankings in Weka. Dataset 2 is contains all the numerical information from the original data and does not contain any information about the rankings from the wine tasters. The aim is to cluster these so that the wines fall into groups similar to the quality attribute of dataset 1, and then create classifiers for this dataset to rate the wine.

Clustering

Dataset 2 requires clustering before it can be classified as the quality attribute of each data point has been removed.

Classification

Naive Bayesian

Naive Bayes classifiers are simple to impliment, fast, and are used in real world situations such as spam filters. They work by looking at the traits of an object, and using each individual trait to determine how likely it is that the object falls into a specific classification. Their downside is that they assume the presence or absense of particular traits has no affect on the classification.

In Weka, we used

Dataset 1

Dataset 2

Support Vector Machine

Support Vector Machine's

Dataset 1

Dataset 2

Neural Network

Dataset 1

Dataset 2

Results

Bibliography

[1] UCI Machine Learning Repository: Wine Quality Data Set. 2014. UCI Machine Learning Repository: Wine Quality Data Set. [ONLINE] Available at: http://archive.ics.uci.edu/ml/datasets/Wine+Quality. [Accessed 01 June 2014].

[2] Weka 3 - Data Mining with Open Source Machine Learning Software in Java . 2014. Weka 3 - Data Mining with Open Source Machine Learning Software in Java . [ONLINE] Available at: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 01 June 2014].

[3] Online CSV to ARFF conversion tool. 2014. Online CSV to ARFF conversion tool. [ONLINE] Available at: http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php. [Accessed 01 June 2014].

[4] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[5] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). "Tackling the poor assumptions of Naive Bayes classifiers". ICML.