Attribute Subset Selection in Physiochemical Based Wine Quality Prediction

Zack Tillotson
Haozhu (Peter) Wang

December 14, 2009

Abstract

We explore a large wine quality data set and search for a "small" subset of attributes which can help consumers confidently predict quality. Twelve classification models are built for every attribute combination. The nearest neighbor models achieve impressive results and a set of five important attributes is found. We decide this set is too large for consumer use but that displaying the "nearest neighbor" wines (similar wines as calculated by using the five important attributes as used in the nearest neighbor models) can lead to helpful consumer results.

Introduction

Wine is growing in popularity, with many people beginning to appreciate it for its complex flavors. A powerful and increasingly important way of helping customers appreciate the wine is via quality assessment and grading. These certifications were traditionally done just using expert analysis, but recently chemical metrics have become an integral part of the process.

As the wine industry grows there are more and more wines available to attempt to find a correlation between the chemical metrics and human quality assessments. This has been attempted using several small wine data sets, but recently a new, high quality wine data set has become available. This set is of Portuguese wines and is of special interest for data mining because of its large size and consistency.

There are twelve chemical measurements of thousands of wines, with the human expert's quality grade as well, with no missing attributes. This allows us to develop an algorithm which attempts to predict the human quality grade given the chemical attributes. A previous effort along these lines was attempted [1] but focused on providing a guide for wine quality certification.

We present an effort to find a way to help the consumer. We would like to find a small subset of the twelve attributes which can predict the quality as well, or nearly as well, as the full set. A small subset would ideally be no more than 3 attributes, or it would become too confusing for a consumer. If this subset does exist it could be printed on each wine bottle as an easy and obvious way of predicting the quality of the wine for the consumer.

Data

The data set is based on a type of Portuguese wine, Vinho Verde. There are approximately 6500 individual wines in the set. Each wine is only represented once in the data set. All chemical tests on the wines were performed by the CVRVV, a Vinho Verde promotional organization [2], this organization also runs the human quality analysis.

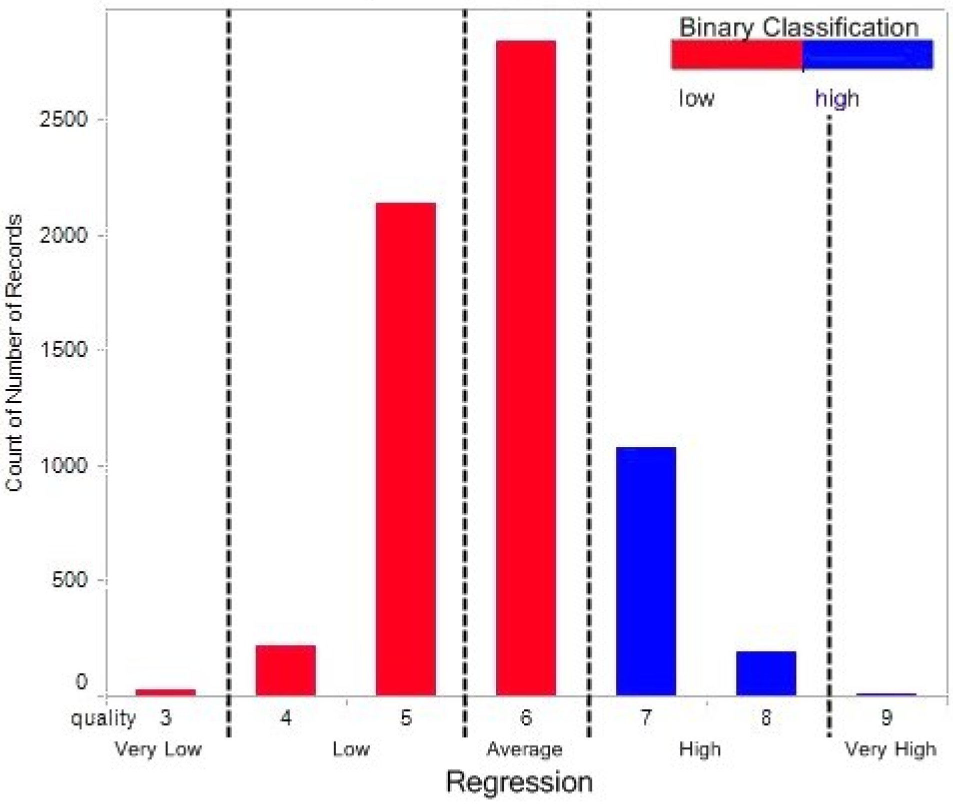
The choice of attributes was mostly logistical. The given twelve attributes are the most common wine metrics, and are available for all wine samples. More obscure chemical tests were not included in the data set in order to ensure no missing attributes were present. There are no non-chemical attributes such as price or vintner, except for one nominal attribute designating the wine as a red or a white.

The chemical measurements were performed by an automated system in order to produce consistent results. For the quality grade, each wine was measured by a minimum of three human experts on a 0-10 scale (0 is bad, 10 is good), with the final grade being the median value. Two methods of analysis were performed - binary classification and regression.

For the binary classification we classified each wine as either "high quality" or "not high quality". In order to be considered “high quality” the wine must score a quality grade of 7 or higher, this is shown in Figure 1.

For regression we discretized the quality into the bins of 0-3, 4-5, 6, 7-8, and 9-10. Because of the small frequency of the 0-3 and 9-10 bins we considered them to be anomalies, this is shown in Figure 1.

Figure 1: Quality Distribution



The way the quality score was generated (taking the median value of 3+ tests) and the high variance in wine quality perception, it is unlikely that wines will be given exceptionally high or low scores. In fact, there are no wines graded as a 1, 2, or 10. There is a relatively large class imbalance, with 80.3% of the wines graded as "not high quality". Besides the red wine/white wine distinction, all other attributes are continuous.

Table 1. Attribute Summary

Attribute	Type	Min	Max	Mean
Fixed Acidity (g/dm3)	Continuous	3.80	15.90	7.22
Volatile Acidity (g/dm3)	Continuous	0.08	1.58	0.34
Citric Acid (g/dm3)	Continuous	0.00	1.66	0.32
Residual Sugar (g/dm3)	Continuous	0.60	65.80	5.44
Chlorides (g/dm3)	Continuous	0.01	0.61	0.06
Free Sulfur Dioxide (mg/dm3)	Continuous	1.00	289.00	30.53
Total Sulfur Dioxide (mg/dm3)	Continuous	6.00	440.00	115.74
Density (g/cm3)	Continuous	0.99	1.04	0.99
pH	Continuous	2.72	4.01	3.22
Sulphates (g/dm3)	Continuous	0.22	2.00	0.53
Alcohol (% by volume)	Continuous	8.00	14.90	10.49
Is Red	Boolean	**True**: 1599 **False**: 4898		
Is High Quality	Boolean	**True**: 1277 **False**: 5220		

Regression Quality	Nominal	**Very Low**: 30 **Low**:2354 **Average**:2836 **High**:1272 **Very High**: 3

Method

The data set contains two distinct types of wine – red and white. Each was thought to be likely to have very different patterns in it with little overlap. The attributes that make for a high quality white wine would make for a very poor red whine, and vice versa. This fact makes the objective more difficult, and other attempts to model the quality on this data set have considered the red and white wines separately. We believe that including this fact models a more realistic consumer experience where the wine being red or white would be just another fact about the wine to take into consideration.

Figure 2 shows the pH vs alcohol content for both red and white wines. It is obvious that the red and white wines are different, but they do seem to share some basic patterns. For example, the higher the alcohol content the higher the quality for both wines. On the other hand, higher quality red wines have a lower pH while high quality white wines are have a wider range.

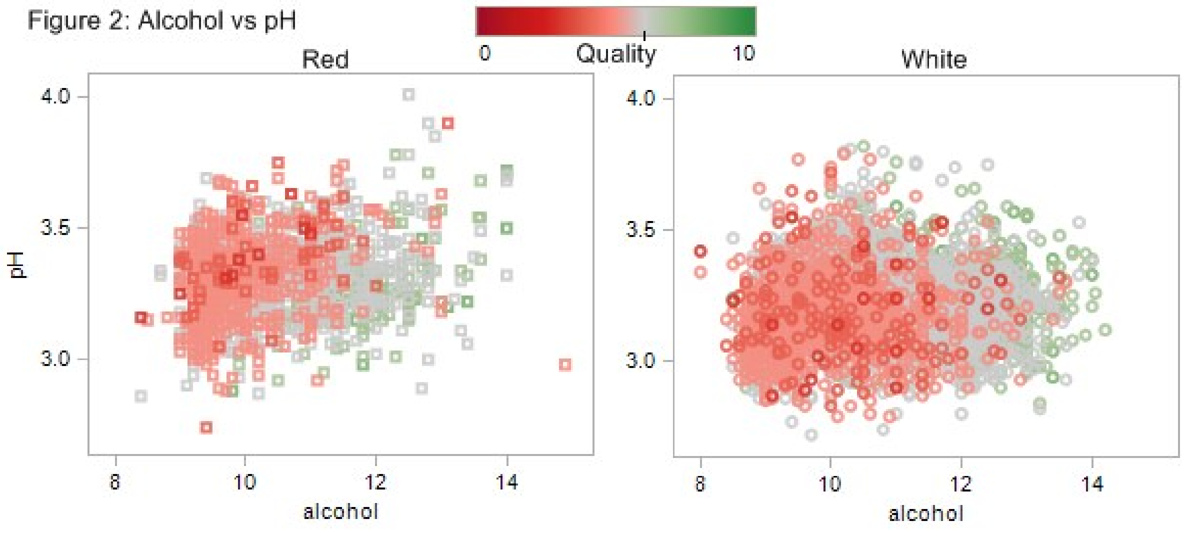


Figure 2: Alcohol vs pH

Using visualizations of the data we found that red wines tend to have less residual sugar, higher pH, and higher fixed and volatile acidity than white wine. To test the hypothesis we ran an analysis which attempted to classify the wines are either red or white, based on their chemical attributes.

A simple ruled based model was able to achieve 93% accuracy

```
(chlorides <= 0.0585) and (total sulfur dioxide > 50.5) => Is Red = False
```

Using the Weka[3] implementation of the J48 algorithm the accuracy of the model was 97.83%. These indicate the hypothesis is correct - the chemical difference between red wine and white wine are very distinct, and we began our attempts to model the quality.

There are many options in what model to choose when classifying the wine qualities. Previous efforts [1] have focused on Support Vector Machines (SVM), and simply repeating their effort was not wanted. Our intent is to use the results of our subset selection to support customers, so it is somewhat preferred that the model we choose is understandable by humans. SVMs are not. Because of these two factors, we decided to avoid SVMs and try decision trees, artificial neural networks, and nearest neighbor algorithms.

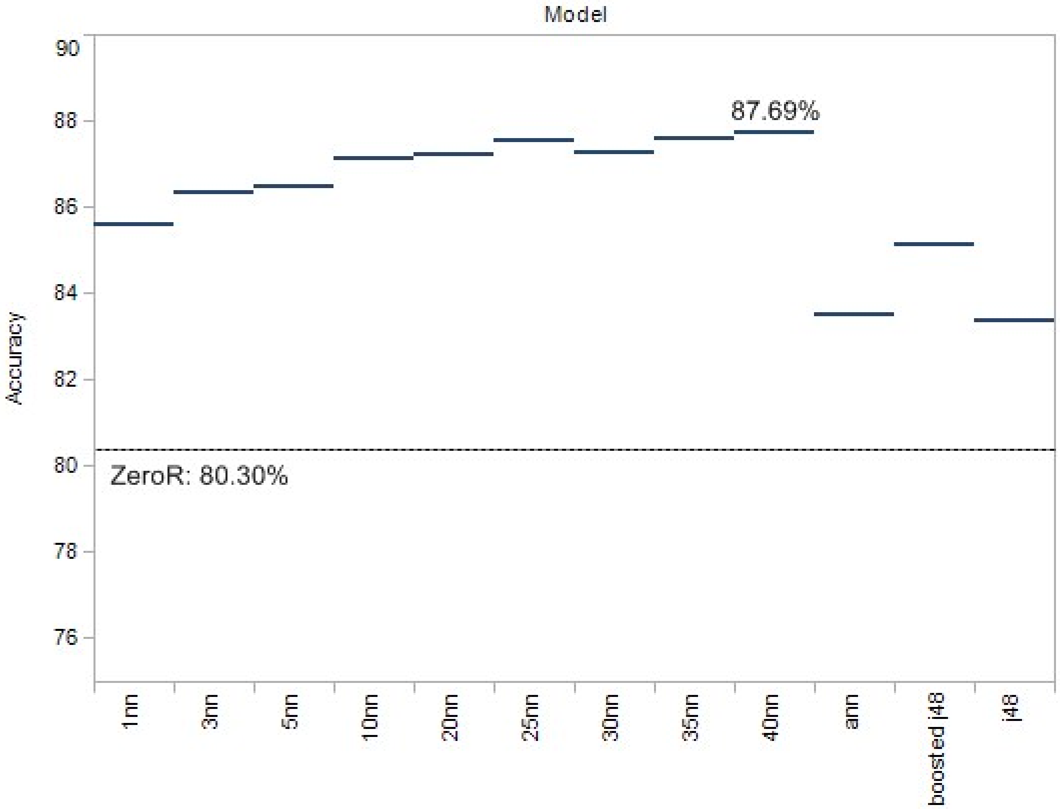
When building a predictive model it is important to consider the quality relative to a baseline. For the classification analysis, the random classifier (which would guess "high quality" or "not high quality" randomly) will be correct 50% of the time since there are two classes. Another, slightly more intelligent model which predicts the higher frequency class ("not high quality") is correct 80.3% of the time, and so this is the model we will use as the lower bound classifier.

We are interested, however, not only in how much better than a baseline our model is, but since we are attempting to find a small subset of attributes which is nearly as good as the full set, we need to establish a baseline full attribute set model. To do this we built several types of models including an Artificial Neural Network (ANN), a simple J48 Decision Tree (J48), a boosted J48 Decision Tree (B-J48), and several Nearest Neighbor models with different numbers of neighbors considered (kNN, where k was 1, 3, 5, 10, 20, 25, 30, 35, or 40).

In order to ensure fair comparisons we used a consistent method to build each model throughout the experiment. Every model is built with the same training set, and tested against the same test set (using a 2:1 split). Obviously when using a subset of the attributes we remove the unneeded attributes from both sets.

Our best baseline model - the 40 nearest neighbor model - had an accuracy of 87.69%. Results for all twelve baseline models for the binary classification analysis are shown in Figure 3.

Figure 3: Baseline Models



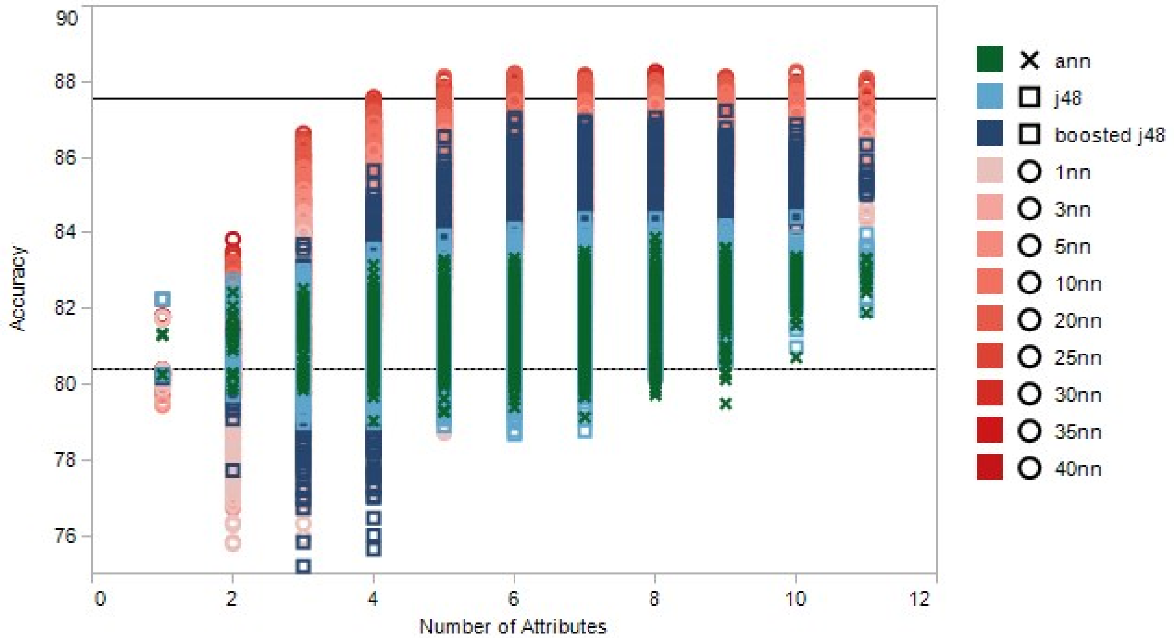
There are many techniques to easily find high quality subsets of attributes. However we decided to iterate over every combination of attributes. The attribute subset selection techniques normally work by first finding the most predictive attribute, then assuming you keep that attribute, finding the next most predictive, and so on until the subset is built. This works well in most situations, but because the relatively small number of attributes allows for an exhaustive search (which guarantees the best results) we went that direction.

There were approximately 4,100 combinations of attribute subsets we tested, and for each we ran the same 12 models as for the baseline. This came out to approximately 49,100 models per analysis.

Binary Classification Results

The full results of our tests are shown in Figure 4. The results are grouped by the number of attributes used to build the model.

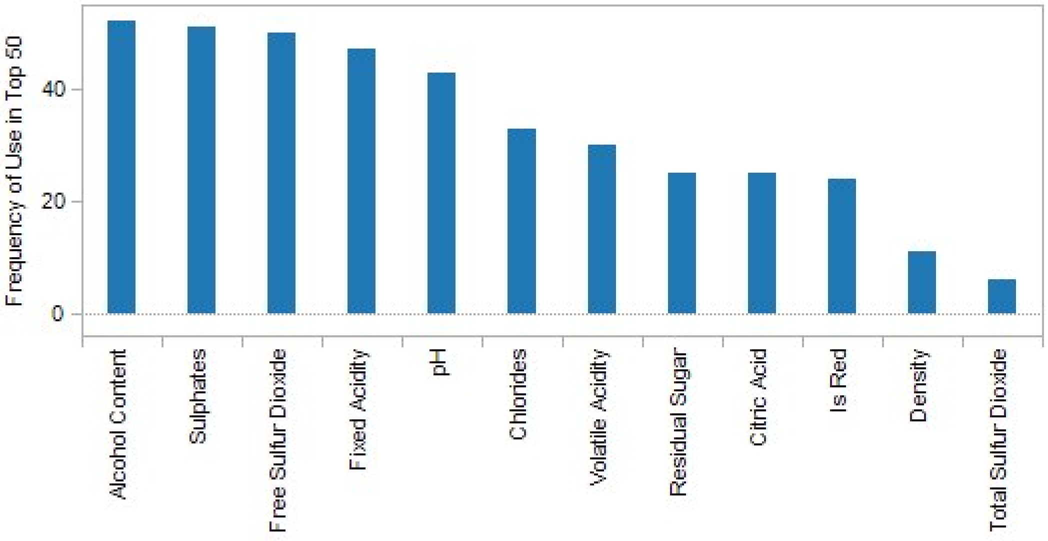
Figure 4: Attribute Subset Accuracy



A maximum accuracy of 88.28% was achieved by a kNN algorithm. It used the 35 nearest neighbors, weighted according to the inverse of the euclidean distance, on 8 attributes - fixed acidity, volatile acidity, citric acid, chlorides, free sulfur dioxide, pH, sulfates, and alcohol content. This performance is a 40% decrease in error over the ZeroR baseline.

This shows an interesting set of results. The main point of interest is that the highest quality models did not increase in quality much between 5 attributes and the max at 10 attributes. This is the type of pattern we were hoping to find in the data. The best 5 attribute combination - Fixed Acidity, Free Sulfur Dioxide, pH, Sulfates, and Alcohol Content - is also the combination of 5 attributes which showed up most often in other top models, of any number of attributes, as shown in Figure 5.

Figure 5: Attribute Frequency In Top Models



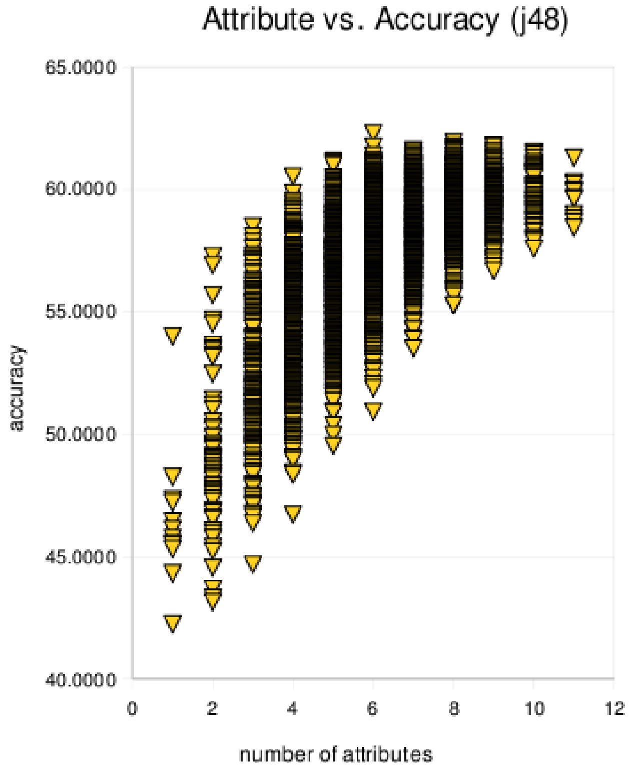
This shows that those 5 attributes are indeed the most important attributes. The other attributes, however, are also important though to differing degrees. Density and Total Sulfur Dioxide, for example, show up only rarely. It would seem at first glance that the relatively low frequency of the "Is Red" attribute is a surprising result. However, the best models were all nearest neighbor algorithms. It stands to reason that for any given wine, the most similar wines would be of the same type and so this attribute is somewhat redundant.

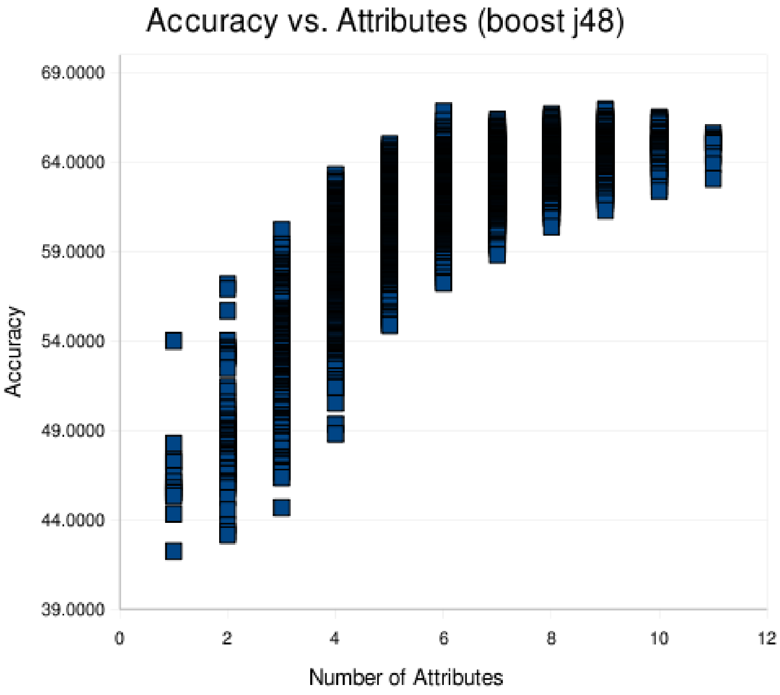
Another surprising result from Figure 4 was that the quality actually decreased beyond 10 attributes. We believe that this is because the non important attributes simply increased the size of the search space without conferring an equal increase in quality correlation. It does seem to allow the models to find a higher minimum though.

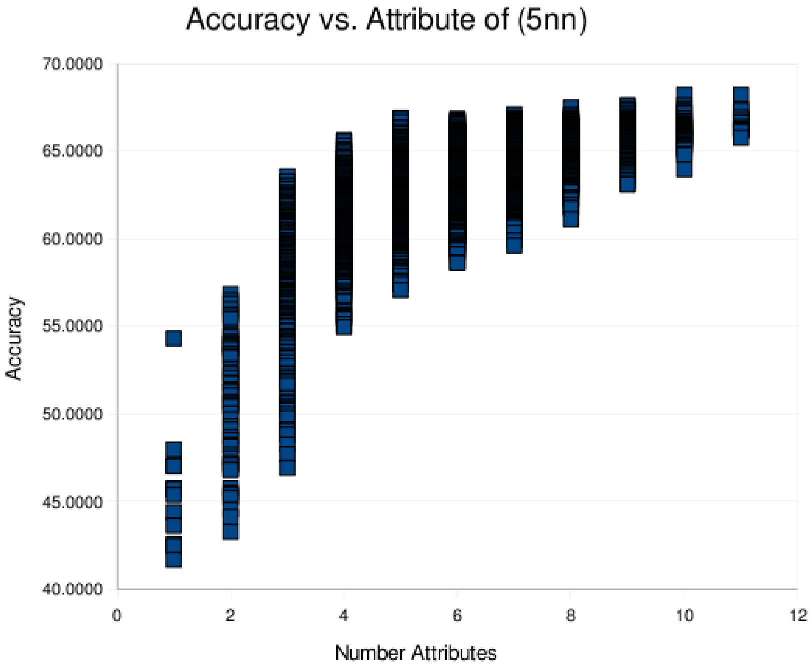
Additionally each model shows the same basic shape, with an increase in max quality up to 4-8 attributes, beyond which the quality actually decreases. The tree algorithms peak at approximately 7 attributes. Using boosting provides a large bonus is accuracy, but does not change the general shape of the curve. The ANN peaks at 8 attributes, we believe this is because ANNs offer a more complex decision boundary, and so can learn the more complicated patterns the higher dimensionality space offers. The nearest neighbor algorithms peak quickly, with a lower number of attributes being needed.

Regression Results

The results of our regression analysis shown a similar accuracy distribution. The following graphs shows the accuracy of various attribute combinations using J48, boosted j48, 5 Nearest Neighbor, and 35 Nearest Neighbor to predict the quality of wine.


Attribute vs. Accuracy (j48)
65.0000
60.0000
55.0000
50.0000
45.0000
40.0000
accuracy
0
2
4
6
8
10
12
number of attributes


Accuracy vs. Attributes (boost j48)
69.0000
64.0000
59.0000
54.0000
49.0000
44.0000
39.0000
Accuracy
0
2
4
6
8
10
12
Number of Attributes

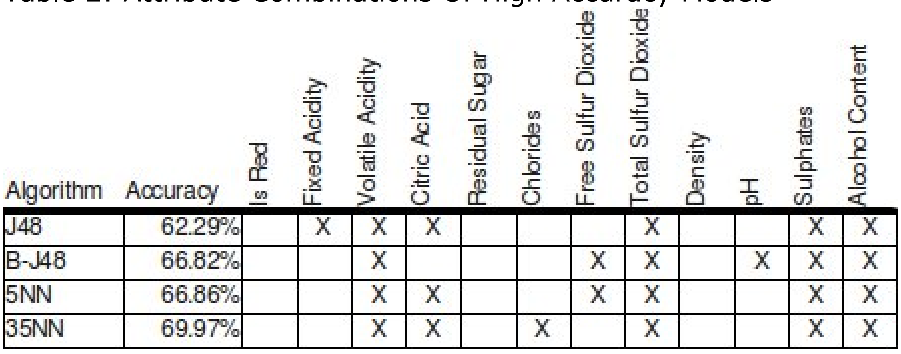


These models exhibit a similar shape to the binary classification models. The accuracies are lower on average, as would be expected with a larger number of bins, but also the variance between using a low number of attributes and a high number increases. The binary classification approach had an accuracy range from low 70s to high 80s, this five bin approach in discretizing the quality has a range from low 40s to low 70s, making discriminatory attributes stand out more.

An interesting pattern is that the accuracy increases in each model with increasing number of attribute combinations up to 6 attributes. Using more attributes that 6 does not significantly improve accuracy. This indicates that 6 attributes are all that is needed to correctly predict the more complex pattern which the five bin classification algorithm requires.

The highest accuracy 6-attribute models for each type of algorithm are shown in Table 2.

Table 2: Attribute Combinations Of High Accuracy Models



Algorithm	Accuracy	Is Red	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol Content
J48	62.29%		X	X	X				X			X	X
B-J48	66.82%			X				X	X		X	X	X
5NN	66.86%			X	X			X	X			X	X
35NN	69.97%			X	X		X		X			X	X

Notice all of these algorithms contain alcohol content, sulfates, some form of total sulfur dioxide, and volatile acidity. Next, we investigated which one attribute would create the best model, the results are shown in

Table3.

Table3: 1 Attribute Model Accuracy

Number of attributes	Attribute type	Algorithm	Accuracy
1	Alcohol Content	35nn	54.50%
1	Citric Acid	35nn	47.90%
1	Density	35nn	47.62%
1	Volatile Acidity	35nn	47.31%
1	Chlorides	35nn	46.72%
1	Is Red	35nn	45.77%
1	Total Sulfur Dioxide	35nn	45.68%
1	Free Sulfur Dioxide	35nn	45.36%
1	Fixed Acidity	35nn	43.96%
1	Sulphates	35nn	42.6437
1	Residual Sugar	35nn	42.10%
1	pH	35nn	41.69%

It is interesting that some attributes which were by themselves generated low accuracy models, like sulphates, are present in the highest accuracy 6 attribute models. The sulphates attribute is inaccurate by itself, but when combined with others it is very accurate.

Conclusion

The main goal of this experiment was to find a small subset of attributes which predict the quality of the wine as well, or nearly as well, as the full attribute set. We feel that 5 attributes found for the binary classification and the 6 attributes found for the regression are perhaps too many to be easily used by the consumer. However, the fact that the nearest neighbor algorithms were the highest quality give us hope. A consumer could be told what quality the nearest k wines, physiochemically speaking, were and could use that to make a decision.

In the future we would like to explore several changes in the experiment, if more time to work on this problem were available. From the data side, we would like to try new attributes. The current attributes were simply the most common tests done on wines, perhaps a lesser used test would have a good correlation with quality. Another option would be to look at non chemical attributes, such as price or vintner of origin. Also, all of the current wines are of the specific Venho Verde type, it would be interesting to expand out into other types of wines.

Bibliography

[1] 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. ISSN: 0167-9236.
Prepublication version available at http://www3.dsi.uminho.pt/pcortez/winequality09.pdf

[2] Comissão de Viticultura da Região dos Vinhos Verdes
http://www.vinhoverde.pt/en/

> The Comissão de Viticultura da Região dos Vinhos Verdes, with the abbreviation of CVRVV, is an inter-professional organization with the objective of representing the interests of the professions involved in the Vinho Verde production and trade and in the defense of the regional and national inheritance of its Denomination of Origin. Legally, this institution is a regional association, a private corporate body of public interest, with undetermined duration.

[3] Waikato Environment for Knowledge Analysis (Weka)
http://www.cs.waikato.ac.nz/ml/weka/