

Predicting Wine's Quality

Introduction

Wine becomes one popular consumption in daily life. The quality of wine is an important element for purchasing selection and produce. To have a clear mind of how to identify the wine quality by provided index would be a helpful technique for customers in selecting process and for producers in producing process. In this project, we present a study for wine quality prediction based on analytical data. What we have done is valuable in business because it can be used to support wine quality evaluation and, in some ways, to improve future wine production's quality.

We propose to use data mining algorithms to predict quality of red and white wine separately. Two large datasets from UCI Machine Learning Repository are used, with red and white samples. Classification, cluster and regression techniques were applied by R, Weka, SAS and Excel.

Dataset Description

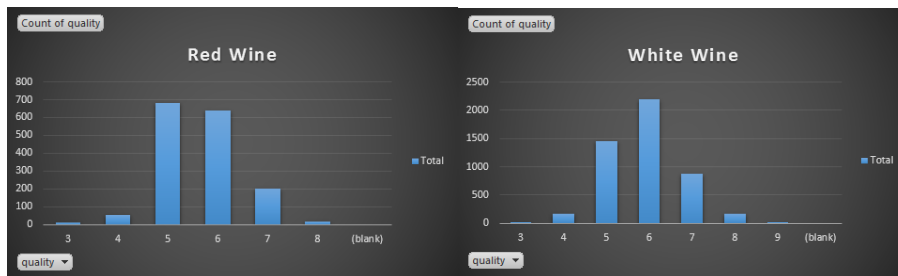
The origin datasets are from UCI, and already have been separated into two, red wine and white wine. The number of attributes in two datasets are same which equals to 12. Attributes are included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality. The input attributes include objective tests and the output is wine grade. Before we do data preprocess, the wine quality are graded between 0 (very bad) to 10 (very excellent). The number of instance for red wine dataset is 1599, and for white wine dataset is 4898.

We do a basic analysis about how many instances are in each quality level in two datasets. The results are following:

Quality/# of instance	Red Wine Dataset	White Wine Dataset
3	10	20
4	53	163
5	681	1457
6	638	2198
7	199	880
8	18	175
9	N/A	5
Total	1599	4898

The original dataset represent quality with numbers (3-9), the distribution of it show as below. As we observe, quality 5 and 6 has the largest instances and the rest of groups has relatively fewer instances. Therefore, we decide to regroup the dataset.

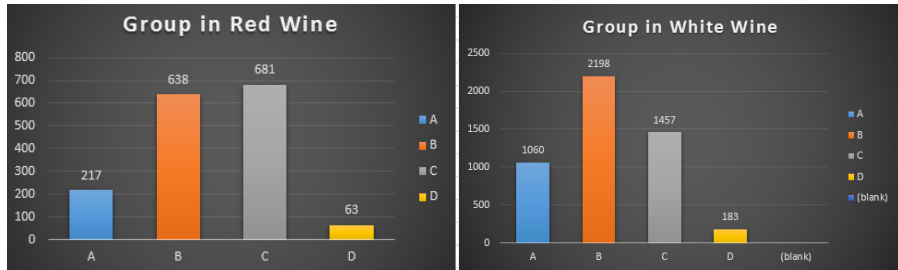
Comment [BY1]: Like it



To have better prediction result, we group the quality level into 4 groups, A(the best), B(good), C(fair), D(the worst). The principle to group is holding the balance of the number of instances in each group. The majority vote of Red Wine is 42.58%(681/1599) and the majority vote of White Wine is 44.87% (2198/4898). Therefore, the quality group is done in following:

Comment [T2]:

The data sets are still unbalanced after grouping. Do you have other considerations on grouping the categories? and what's the mapping between original categories and regrouped ones?



Data Preprocess

Based on our knowledge and intuition, we think the qualities of red wine and white wine cannot be compared with each other. The level of quality 10 of red wine is not as same as the level of quality 10 of white wine. Different variable value may contribute differently to the final quality result because they might have different types of association with quality. Red wine generally speaking contains more sugar than white wine, thus the variable of residual sugar is different. Based on this perception, it is not valuable to combine two datasets and compare with another wine's quality. Therefore, we still keep the red wine dataset and white wine dataset separately.

As mentioned above, we transformed the quality from 0-10 to A, B, C and D. We transform the "quality" to ordinal type in R. Shown as below:

```
10 red$quality=ordered(red$quality)
11 white$quality=ordered(white$quality)
12 str(red)
13 str(white)
$ quality : Ord.factor w/ 4 levels "A"<"B"<"C"<"D": 2 2 2 2 2 2 2 2 2 2 ...
```


Then, we calculate the info gain of each dataset. All the variables shown below.

```
> InfoGainAttributeEval(quality ~ . , data = red)
fixed.acidity    volatile.acidity    citric.acid
0.02159545      0.12822396             0.07744895
residual.sugar   chlorides      free.sulfur.dioxide
0.00000000      0.03635079             0.00000000
total.sulfur.dioxide    density    pH
0.07032455             0.05698079    0.00000000
sulphates              alcohol    group
0.11935661             0.23239330    1.70906164
```

As we can see, from the result that “residual.sugar”, “free.sulfur.dioxide” and “pH” have zero info gain in red wine dataset, thus we remove them all. Meanwhile, we also remove “fixed.acidity” and “citric.acid” which has very low info gain and only keep “volatile acidity” in term of acidity measurement to reduce information crowding. The rest of the variables are all included in the further data analysis.

Red wine prediction:

fixed.acidity		
volatile acidity		
citric.acid		
residual.sugar		
Chlorides		
free.sulfur.dioxide		
total sulfur dioxide		
Density		
pH		
Sulphates		
Alcohol		




quality

```
> InfoGainAttributeEval(quality ~ . , data = white)
fixed.acidity    volatile.acidity    citric.acid
0.01247295      0.05710199             0.06720771
residual.sugar   chlorides      free.sulfur.dioxide
0.06183840      0.07743839             0.05966286
total.sulfur.dioxide    density    pH
0.06286092      0.11773520             0.01619921
sulphates          alcohol    group
0.01477956        0.20783658             1.86171493
```

Similarly based on white wine info gain, we remove the unassociated variables to prevent information crowding and improve the prediction efficiency and accuracy.

White Wine Quality Prediction:

fixed.acidity		
volatile acidity		
citric.acid		
residual.sugar		
chlorides		
free.sulfur.dioxide		
total.sulfur.dioxide		
density		
pH		
sulphates		
alcohol		



quality

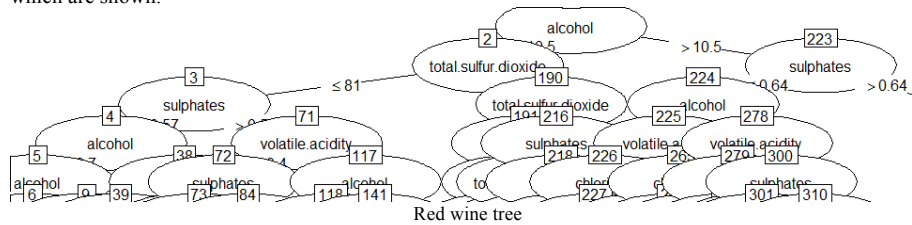
Analysis Experiments

Decision Tree:

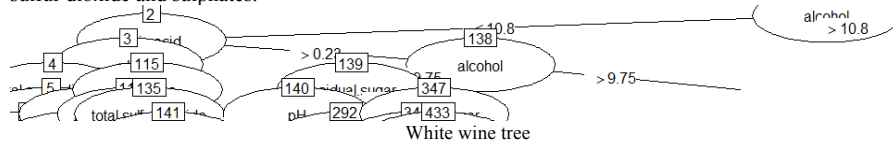
We use the filtered variables to build Decision Tree model in both R and Weka and test the result via cross-validation method. Sample results are shown below.

In R:

We load the data, install the package and run the R to plot the tree for both red wine and white wine, which are shown.



After we plot the tree in R, we can see the first node is alcohol, other important nodes also include total sulfur dioxide and sulphates.



For white wine, the first node is alcohol as well, the next one are citric acid and residual sugar which is significantly different from red wine.

In Weka:

```
Correctly Classified Instances      964      60.2877 %
Incorrectly Classified Instances    635      39.7123 %
Kappa statistic                    0.3755
Mean absolute error                 0.2162
Root mean squared error             0.4007
Relative absolute error             67.5918 %
Root relative squared error         100.2165 %
Total Number of Instances          1599

=== Detailed Accuracy By Class ===

    TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
    0.734    0.282    0.659    0.734    0.694    0.761    C
    0.525    0.238    0.594    0.525    0.557    0.673    B
    0.571    0.083    0.519    0.571    0.544    0.789    A
    0.079    0.021    0.135    0.079    0.1    0.585    D
Weighted Avg.   0.603    0.227    0.593    0.603    0.596    0.723

=== Confusion Matrix ===

  a  b  c  d  <-- classified as
500 142  20  19 | a = C
201 335  92  10 | b = B
 17  73 124   3 | c = A
 41  14   3   5 | d = D
```

M1 result for red wine

PARAMETERS	U <use unpruned tree>	C <pruning confidence>	M <minimum number of instances>	numFolds	B <Use binary splits only>	S <Do not perform subtree raising>	Accuracy% in Weka (red/white)

Comment [T3]:
The splitting nodes could be varied between decision tree models, so which model does this tree represent?

M1	FALSE	0.5	2	5	FALSE	FALSE	60.28/ 58.17
M2	FALSE	0.5	2	10	FALSE	FALSE	60.22/ 60.04
M3	FALSE	0.25	2	5	FALSE	FALSE	60.35/ 58.30
M4	FALSE	0.75	2	5	FALSE	FALSE	59.91/ 58.16
M5	FALSE	0.5	3	5	FALSE	FALSE	60.28/ 56.92

As we tune the parameters, we find out for red wine the highest accuracy is M3 and for white wine the highest accuracy is M2.

Naive Bayes:

The standard Naïve Bayes classifier computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule. Two versions of rules are applied, one is to assume they are following normal distribution; the other is to apply discretization to numeric variables first and then use Naïve Bayes algorithm to predict results since the variables may not follow normal distribution in the whole population.

The test method we choose is the same as previous one -- cross validation and the number of folds is 5. The accuracy before discretization and after discretization is slightly different, which are 57.66% and 59.10% for red wine and 45.81% and 49.42% for white wine, thus after discretization accuracy is higher than the one before discretization.

After we run code in R of dependent variables' conditional probability, we find out that for red wine, both alcohol and total sulfur dioxide has strong association with quality. For white wine, alcohol and citric acid has stronger association with quality result than other variables.

Algorithm	Red wine (%)	White wine (%)
NB	57.66	45.81
NB with discretization	59.10	49.42

```

Correctly Classified Instances      2421          49.4283 %
Incorrectly Classified Instances    2477          50.5717 %
Kappa statistic                    0.2505
Mean absolute error                 0.2749
Root mean squared error             0.4031
Relative absolute error             83.0416 %
Root relative squared error         99.0931 %
Total Number of Instances          4898

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.388    0.293    0.519    0.388    0.444    0.592    B
      0.648    0.286    0.49    0.648    0.558    0.751    C
      0.583    0.177    0.476    0.583    0.524    0.797    A
      0.033    0.005    0.207    0.033    0.057    0.743    D
Weighted Avg.   0.494    0.255    0.489    0.494    0.481    0.689

=== Confusion Matrix ===
      a  b  c  d  <-- classified as
853 765 571  9 |  a = B
415 944  85 13 |  b = C
308 133 618  1 |  c = A
 68  86  23  6 |  d = D

```

White wine after discretization

SVM

Support Vector Machines is the algorithm seeking for optimum separating hyperplane between the two classes by maximizing the margin between the classes' closest points.

I started training using the default settings in algorithm parameters, which uses "PolyKernel" as the kernel used in training and predicting, 1 as the cost of misclassification (c value), -1 as number of folds for cross-validation, random number seed of 1. The red wine result gives me accuracy of 58.10% returned by Weka shown as below. Then I tune c value to 10, the result is slightly higher 58.14%. When c value reaches 100, the accuracy result is 58.53%. However, for white wine, the model with highest accuracy is when c=0.5

Comment [BY4]: Just include the evaluation numbers that you used for your analysis

```
Correctly Classified Instances      929          58.0988 %
Incorrectly Classified Instances    670          41.9012 %
Kappa statistic                    0.2861
Mean absolute error                 0.295
Root mean squared error             0.3772
Relative absolute error             92.2261 %
Root relative squared error         94.3306 %
Total Number of Instances         1599

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.783    0.33    0.638    0.783    0.703    0.747    C
      0.621    0.382    0.519    0.621    0.565    0.619    B
      0         0         0         0         0         0.821    A
      0         0         0         0         0         0.651    D
Weighted Avg.   0.581    0.293    0.479    0.581    0.525    0.702

=== Confusion Matrix ===

  a  b  c  d  <-- classified as
533 148  0  0 |  a = C
242 396  0  0 |  b = B
 12 205  0  0 |  c = A
 49  14  0  0 |  d = D
```

SVM (c=1) for red wine

Algorithm	Red Wine Accuracy	White Wine Accuracy
SVM (c=1)	58.10	49.67
SVM (c=0.5)	58.14	49.78
SVM (c=10)	58.53	49.51

Generate Linear Multiple Regression Model

In this part, we want to find the regression relationship between quality and selected attributes, volatile acidity, chlorides, total sulfur dioxide, density, sulfates and alcohol. In order to get the more accurate regression equation, we use number value of quality level from original dataset as dependent variable, instead of the grade level of quality. The value of quality is the number level from 3 to 8 in red wine dataset/3 to 9 in white wine dataset (the worst to the best).

Red wine regression model:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-3.16245	10.28803	-0.31	0.7586
volatile_acidity	1	-1.13968	0.09703	-11.75	<.0001
chlorides	1	-1.70995	0.39186	-4.36	<.0001
total_sulfur_dioxide	1	-0.00229	0.00050893	-4.51	<.0001
density	1	6.13471	10.23158	0.60	0.5489
sulphates	1	0.90236	0.11224	8.04	<.0001
alcohol	1	0.28280	0.01903	14.86	<.0001

Regression equation for red wine quality:

$$\text{Red Wine Quality} = -3.162 - 1.139va - 1.710c - 0.002tsd + 6.135d + 0.902s + 0.283a$$

(va = volatile acidity, c = chlorides, tsd = total sulfur dioxide, d = density, s = sulphates, a = alcohol)

From the parameter estimates result, it is clear that volatile acidity, chlorides and total sulfur dioxide have negative relationships with red wine quality. But the density, sulfates and alcohol have positive relationships with red wine quality. According to the P value, most of variables, except intercept and density, have a small enough P value, which means the experiment needs to reject H0 assumption and accept H1 assumption. Therefore, these independent variables have linear relationships to red wine quality.

Comment [T5]:
Which measurements do you use to evaluate the linear model's performance?

White wine regression model:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	119.35487	13.12514	9.09	<.0001
citric_acid	1	0.26031	0.09475	2.75	0.0060
residual_sugar	1	0.06300	0.00529	11.91	<.0001
chlorides	1	-1.87644	0.55489	-3.38	0.0007
total_sulfur_dioxide	1	0.00017101	0.00031384	0.54	0.5859
density	1	-117.09188	13.09882	-8.94	<.0001
sulphates	1	0.73807	0.10231	7.21	<.0001
alcohol	1	0.20226	0.01890	10.70	<.0001

Regression equation for white wine quality:

$$\text{White Wine Quality} = 119.355 + 0.260ca + 0.063rs - 1.876c + 0.00017tsd - 117.092d + 0.738s + 0.202a$$

(ca = citric acidity, rs = residual sugar, c = chlorides, tsd = total sulfur dioxide, d = density, s = sulphates, a = alcohol)

Similarly, to the regression analysis in red wine part, citric acid, residual sugar, total sulfur dioxide, sulfates and alcohol have positive relationships with white wine quality. And the chlorides and density

have negative relationships with white wine quality. Most of P values, except total sulfur dioxide's, are small enough to support linear regression assumption.

Relationships Overview:

Variables	Relationship with red wine quality	Relationship with white wine quality
total sulfur dioxide	(-)	(+)
density	(+)	(-)
sulphates	(+)	(+)
alcohol	(+)	(+)
volatile acidity	(-) red wine only	N/A
citric acidity	N/A	(+) white wine only
residual sugar	N/A	(+) white wine only

(+: positive, -: negative)

When we review the estimate parameters of regression equations, total sulfur dioxide and density have absolutely reverse relationships with two wines' qualities. And based on the parameter value of density, density is an important factor to affect the quality of wine. In red wine, parameter of density is 6.137 and in white wine, parameter of density is -117.092. Especially in white wine regression model, the P value of density is smaller than 0.0001, which means it is a decisive part participating into linear regression model. Therefore, density is a crucial element to decide the quality of white wine.

Conclusion

All of the test results for both red wine and white wine are higher than the majority vote which proves that the models are valuable in predicting the red and white wine quality. The impact of each variable for red wine and white wine vary. However, we can see the alcohol has the most significance in prediction wine quality. For red wine, residual sugar has no info gain so that we haven't even included into our analysis, however, it impacts the white wine quality greatly.

According to regression results, total sulfur dioxide and volatile acidity have negative relationships with red wine's quality. And density, sulphates, alcohol have positive relationships with red wine's quality. For white wine, only density has a negative relationship with quality, other selected variables, total sulfur dioxide, sulfates, alcohol, citric acidity and residual sugar all have positive relationships with white wine's quality. Most variables are supportive to linear relationships with quality.

Comment [T6]:
Data Collection: Proficient
Analysis Method: Proficient
Interpretation: Proficient
Conclusion: Proficient