Prediction of wine quality from physicochemical properties

Michal Horák

***Abstract*— With data mining techniques we can predict wine taste preferences based on physicochemical properties from wine analyses. This work solves the regression problem using regression tree algorithm. Setting splitmin value to 85 is possible to find prediction model with over 88% accuracy. This model can be used for wine evaluation or for improving wine production.**

I. ASSIGNMENT

Use the regression tree algorithm to find best prediction model. Make any necessary experiments which clearly demonstrate how to set up chosen algorithm. Chose minimize count of data for next node splitting.

II. INTRODUCTION

The right wine quality evaluation is important for the market. Evaluation prevents the illegal adulteration of wines and assures quality for the wine market.

Wine can be classified by human experts or by physicochemical laboratory tests – pH or alcohol values or density determination.

But there is another method how to classify wine. We can use Data mining techniques on collected wine sets and evaluate wine quality by given raw data.

The dataset is in (.cvs) format and represent ***white*** wine set with over than 4000 rows and each row has 11 attributes. Table 1 present basic summarization of set. Each sample was also evaluated by mark in a range from 0 (very bad) to 10 (excellent) which represent a wine quality.

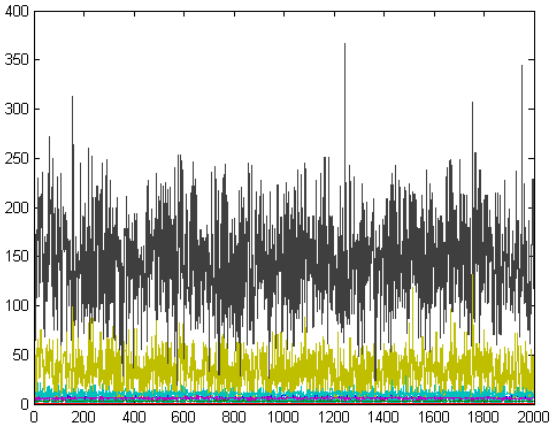

Fig. 1. Wine sets which isn't normalized. And range is between values 0 to 366.

TABLE I
BASIC SUMMARIZATION OF WHITE WINE SET

Attribute (units)	min	mean	max
Fixed acidity (g(tartaric acid)/dm3)	5	7	14,2
Volatile acidity (g(acetic acid)/dm3)	0,08	0,26	1,005
Citric acid (g/dm3)	0,01	0,34	1,66
Residual sugar (g=dm3)	0,6	5	31,6
Chlorides (g(sodium chloride)/dm3)	0,017	0,044	0,346
Free sulfur dioxide (g/dm3)	3	34	146,5
Total sulfur dioxide (g/dm3)	19	144	366,5
Density (g/dm3)	0,9881	0,9944	1,0103
PH	2,72	3,25	3,9
Sulphates (g(potassium sulphate)/dm3)	0,25	0,465	0,97
Alcohol (% vol.)	8,5	10	14

III. METHODOLOGY

A. *Setup of experiments*

For this work is used data mining technique which solves the regression problems. Also is used regression tree algorithm. Experiments I done is about to find best ***splitmin*** value - minimize count of data for next node splitting. Second important setting is to find right ratio of training data to testing data. And finally I did some experiments with nFolds which is parameter for cross validation algorithm.

B. *Configuration of algorithms*

From my observation I set following parameters to these values:

- **Training Fraction** parameter I set to 0.7. So the wine set is divided to 70% for training data and 30% for testing data.
- Value of parameter **splitmin** I set to 85.
- And **nFolds** parameter I set to a value 10.

C. *Tool used*

For this semestral work was used mathematical tool called Matlab - version 7.9.0.529 (R2009b). For my experiments I used prepared scripts which I modified for my needs for

experiments. These scripts were prepared by our data mining lectors. Also these scripts use statistical functions in Matlab.

IV. Experiments

Dataset has all attributes in numerical interpretation. But there were a few rows, with zero attributes, so I had to delete that rows. For experiments I chose 2000 rows for quicker processing.

Wine set wasn't normalized, so there was a range of values from 0, 0100 to 366,500. This fact is shown in Figure 1. Because that range, the first step was to normalize the data to range from 0 to 1 value. Normalized wine set is shown in figure 2, where you can see, all attributes has the range between 0 and 1.

After normalization the next step was to divide wine set to training fraction and test fraction. For this work 70% data (1333 records) goes for training phase and remaining 30% (667 records) is used for testing.

Next step was to set the splitmin. I discovered the 85 is an optimum value for that parameter. This fact you can see at figure No. 3.

After dividing wine set to two parts and setting splitmin, the cross validation (CV), which can be used to compare the

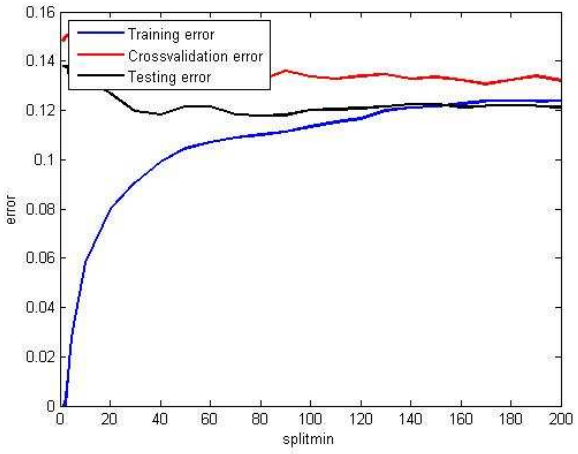

Fig. 3. Splitmin parameter was set according to my observation and testing the ratio of splitmin level to testing error.

performances of different predictive modeling procedures, was used. [1] There I set the nFolds which is a parameter for CV function. I created my own script which for iFold to nFolds, where iFold increasing by +1, computes the CV error. After ending the iteration, the iFold, where had been the lowest CV error, is picked and then is applied for creating better prediction model.

After processing all given algorithms the prediction I've achieved was **88,29 %** for testing subset.

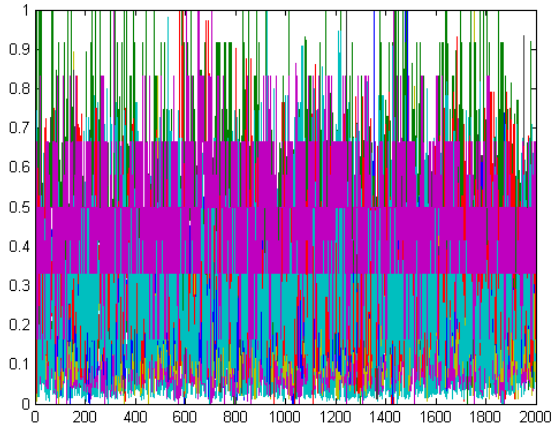

Fig. 2. This figure shows a normalized wine set from range 0 to 1.

V. Discussion

The result is in my humble opinion good. Paulo Cortez (Ph.D. in Computer Science) who is assistant professor at University of Minho, where he develops teaching and R&D activities. [2] He did the data mining research with the same dataset. And in his work he obtained a global accuracy of 89.0% (red) and 86.8% (white). So I have found better prediction model for white wine set.

VI. Conclusion

Due to advances in the data mining, we can extract knowledge from raw data. In this work I solved the regression problem. The algorithm I chose is called regression tree. With this algorithm and with right setting of parameters like a ratio of training and testing data, the splitmin value and nFolds, we can achieve very good results.

The prediction model in this work attacks almost 90% board line which is quite good result.

References

[1] G. Williams, Data mining, [online], [cit. 14. 4. 2010], Available from WWW: http://www.togaware.com/datamining/survivor/Cross_Validation.htm

[2] 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.

[3] Wine Classification, online], [cit. 14. 4. 2010], Available from WWW: http://www.metalimagination.com/wineclassification.html