A DATA SCIENCE STUDY FOR DETERMINING FOOD QUALITY: AN APPLICATION TO WINE

A.E. OZALP AND I.N. ASKERZADE

ABSTRACT. In this paper, wine quality is investigated based on physicochemical ingredients which include fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate and alcohol, by ANFIS (Adaptive Neuro Fuzzy Inference System) method and by random forest algorithm which is a powerful classification algorithm. Although this study specifically investigate the relation between physicochemical ingredients and the quality of wine, the results can be adapted to determination of the quality of any food product in terms of the ingredients.

1. INTRODUCTION

Producers are in a competition of making their production more competitive, since human’s existential pleasure is depended mostly on consumption. This competition exists also in the food industry which is fundamental for human life. It is difficult to predict the quality of the food or to measure the consumer’s tasting. Fuzzy logic is an effective method for complex concepts like human’s emotions and feelings to reduce the complexity. According to the archeological records, wine has been in human’s life for 9500 years [1]. Like the other foods and drinks, quality of wine also has been an important factor for people’s preference of wine consumption. Although determining and controlling the quality of wine seems oenologists’ and food engineers’ field, in recent years, with the developments of data mining, data scientists start to contribute in the area of the quality control.

In this study, to decide the quality of red wine, a fuzzy logic method which is ANFIS and the random forest algorithm will be used. Although the concept of quality is heuristic and includes emotion, fuzzy logic methods are supposed to reduce the heuristic complexity. As another method for quality control, random forest algorithm is chosen, because it is the most successful classification algorithm according to Delgado [2]. Quality control will be done using UCI (University of...
California, Irvine) machine learning repository’s red wine quality dataset which is uploaded by Paulo Cortez (a website which has approximately 400 data sets) [13]. There are two dataset for red and white wine in the repository. In this paper, red wine quality is investigated.

In literature, the first study for food quality assessment using fuzzy logic is in 1995 [3]. Goel and Perrot give a review of previous research on food quality using fuzzy logic [4, 5]. In [6] and [7] fuzzy logic methods are used to classify the data sets. The mathematical basis of fuzzy logic is also developed in [8].

2. DATASET

UCI machine learning repository’s red wine quality dataset is used to determine the quality of red wine. In the creation of the dataset, the physicochemical properties are given for 1599 red wines for which a number of Oenologists taste the wines and give a score between 0, to the worst, and 10, to the best, to each wine. The quality value of each wine is determined from the average of the scores given by the Oenologists for that wine. Inputs of the dataset include the quality value and the amount of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates and alcohol. The aim of this study is to determine the quality of red wine by using the fuzzy logic and random forest algorithm for the physicochemical properties of red wine, and compare the results by the scores of Oenologists’.

Table 1 shows first ten data of wine quality data set. Table 2 [9] gives the statistics of wine quality data set.

Table 1. First ten data of data set.

<table>
<thead>
<tr>
<th>fixed acidity</th>
<th>volatile acidity</th>
<th>citric acid</th>
<th>residual sugar</th>
<th>chlorides</th>
<th>free sulfur dioxide</th>
<th>total sulfur dioxide</th>
<th>density</th>
<th>pH</th>
<th>sulphates</th>
<th>alcohol</th>
<th>quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.4</td>
<td>0.50</td>
<td>1.9</td>
<td>96</td>
<td>11</td>
<td>51</td>
<td>0.0374</td>
<td>3.51</td>
<td>0.58</td>
<td>9.4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7.8</td>
<td>0.88</td>
<td>2.6</td>
<td>98</td>
<td>25</td>
<td>67</td>
<td>0.9968</td>
<td>3.20</td>
<td>0.68</td>
<td>9.8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7.8</td>
<td>0.76</td>
<td>2.3</td>
<td>92</td>
<td>15</td>
<td>54</td>
<td>0.9970</td>
<td>3.28</td>
<td>0.65</td>
<td>9.8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>11.2</td>
<td>0.56</td>
<td>1.9</td>
<td>75</td>
<td>17</td>
<td>60</td>
<td>0.9980</td>
<td>3.16</td>
<td>0.58</td>
<td>9.8</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7.4</td>
<td>0.66</td>
<td>1.8</td>
<td>75</td>
<td>13</td>
<td>40</td>
<td>0.9978</td>
<td>3.51</td>
<td>0.56</td>
<td>9.4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7.9</td>
<td>0.86</td>
<td>1.6</td>
<td>69</td>
<td>15</td>
<td>59</td>
<td>0.9984</td>
<td>3.39</td>
<td>0.46</td>
<td>9.4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7.3</td>
<td>0.65</td>
<td>1.2</td>
<td>65</td>
<td>15</td>
<td>21</td>
<td>0.9968</td>
<td>3.39</td>
<td>0.47</td>
<td>10.0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7.8</td>
<td>0.58</td>
<td>2.0</td>
<td>73</td>
<td>9</td>
<td>18</td>
<td>0.9968</td>
<td>3.36</td>
<td>0.57</td>
<td>9.5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7.5</td>
<td>0.56</td>
<td>6.1</td>
<td>71</td>
<td>17</td>
<td>102</td>
<td>0.9978</td>
<td>3.35</td>
<td>0.40</td>
<td>10.5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2. Statistics of data set.

<table>
<thead>
<tr>
<th>Attribute(s)</th>
<th>Mean</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed acidity(g(tartaric acid)/dm³)</td>
<td>4.6</td>
<td>15.9</td>
<td>1.3</td>
</tr>
<tr>
<td>Volatile acidity(g(acetic acid)/dm³)</td>
<td>0.1</td>
<td>1.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Citric acid(g/dm³)</td>
<td>0.8</td>
<td>1.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Residual sugar(g/dm³)</td>
<td>0.9</td>
<td>15.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Chlorides(g(sodium chloride)/dm³)</td>
<td>0.01</td>
<td>0.61</td>
<td>0.08</td>
</tr>
<tr>
<td>Free sulfur dioxide(mg/dm³)</td>
<td>1</td>
<td>72</td>
<td>14</td>
</tr>
<tr>
<td>Total sulfur dioxide(mg/dm³)</td>
<td>6</td>
<td>280</td>
<td>46</td>
</tr>
<tr>
<td>Density(g/cm³)</td>
<td>0.990</td>
<td>1.004</td>
<td>0.996</td>
</tr>
<tr>
<td>pH</td>
<td>2.7</td>
<td>4.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Sulphates(g(potassium sulphate)/dm³)</td>
<td>0.3</td>
<td>2.0</td>
<td>0.7</td>
</tr>
<tr>
<td>Alcohol(vol%)</td>
<td>8.4</td>
<td>14.9</td>
<td>10.4</td>
</tr>
</tbody>
</table>
2.1. **Classification.** Decision tree is one of the most popular classification method in data mining. This classifier is a supervised learning method and it is built from a set of training examples. A decision tree has nodes, branches and leaves. The name of the top node is root node. All decision trees start from the root node. They grow top to bottom at each level. Nodes are connected by branches and leaves [10].

![Figure 1. An example of decision tree.](image)

Another algorithm for classification is Random Forest Algorithm. As it is understood from its name, random forest algorithm creates the forest with a number of trees. The random forest algorithm is developed by Leo Breiman and Adele Cutler [2]. This approach is a supervised, decision tree based algorithm. It is the most successful classification algorithm [2].

ANFIS is a network structure which based on Takagi–Sugeno fuzzy inference system. It suggested by Roger Jang in 1993 [11]. This system takes advantages of fuzzy logic’s decision making feature and artificial neural networks’ learning ability, and thus, it combines human’s intelligence and artificial intelligence.

2.2. **ANFIS Architecture.** ANFIS has 5 layers which is shown in Figure 2.

**A.:** Layer 1: This layer is fuzzification layer.

\[
O_i^1 = \mu_{A_i}(x), \quad i = 1, 2
\]

\[
O_i^2 = \mu_{B_i}(y), \quad i = 1, 2
\]

where \(x\) and \(y\) are input values, \(A_i\) and \(B_i\) are linguistic variables. \(A_i\) and \(B_i\) are degrees of membership of membership functions. For membership degree calculation, Gaussian and Bell membership functions are used.

**B.:** Layer 2: This layer is called the rule layer.

\[
w_i = \mu_{A_i}(x) \times \mu_{B_i}(y), \quad i = 1, 2.
\]

**C.:** Layer 3: Third layer is the normalization layer.
Figure 2. ANFIS architecture.

\[ \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2. \]

D.: Layer 4. In this layer defuzzification is done.

\[ O^3_i = \bar{w}_i f_i = \bar{w}_i (p_i(x) + q_i(y) + r_i), \quad i = 1, 2. \]

In this formula \( w_i \) is layers' output value, \( \{p_i, q_i, r_i\} \) is the set of fuzzy inference system's posteriori parameters.

E.: Layer 5. This layer is called output layer.

\[ O^4_i = y = \sum_i \bar{w}_i f_i = \sum_i \frac{w_i f_i}{w_i f_i}, \quad i = 1, 2. \]

2.3. Hybrid Learning Algorithm. There are two parameter sets; \( S_1 \) states input parameter set \( S_2 \) states output parameter set. Total parameter set is \( S = S_1 + S_2 \).

There are two states in hybrid learning algorithm which are backfeed and forward-feed. \( S_1 \) is obtained by least squares method and \( S_2 \) is obtained by linearization [12]. Output is as follows:

\[
\begin{align*}
f &= \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2 \\
f &= w_1 f_1 + w_2 f_2 \\
f &= (w_1 x) p_1 + (w_1 y) q_1 + (w_1) r_1 + (w_2 x) p_2 + (w_2 y) q_2 + (w_2) r_2
\end{align*}
\]

3. APPLICATIONS AND RESULTS

A. Random forest application in WEKA (Waikato Environment for Knowledge Analysis)

To make the classification more clear in data set, data which have output value 3 and 4 will be classified as low, 5 as medium and 6,7,8 as high.
60% of data is used for training, and remaining 40% is used for testing the data. Obtained result are presented in Figure 4.

--- Run information ---
Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation: wine
Instances: 1599
Attributes: 12
- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality
Test mode: split 60.0% train, remainder test

--- Classifier model (full training set) ---
RandomForest

--- Summary ---
Correctly Classified Instances 493 77.0313 %
Kappa statistic 0.5512
Mean absolute error 0.2229
Root mean squared error 0.329
Relative absolute error 62.8931 %
Root relative squared error 78.2819 %
Total Number of Instances 640

--- Detailed Accuracy By Class ---
<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,000</td>
<td>0,000</td>
<td>0,000</td>
<td>0,000</td>
<td>0,000</td>
<td>0,000</td>
<td>0,000</td>
<td>0,789</td>
<td>0,148</td>
<td>low</td>
</tr>
<tr>
<td>0,747</td>
<td>0,192</td>
<td>0,758</td>
<td>0,747</td>
<td>0,753</td>
<td>0,557</td>
<td>0,866</td>
<td>0,831</td>
<td>0,876</td>
<td>medium</td>
</tr>
<tr>
<td>0,838</td>
<td>0,258</td>
<td>0,780</td>
<td>0,838</td>
<td>0,888</td>
<td>0,584</td>
<td>0,875</td>
<td>0,876</td>
<td>0,832</td>
<td>high</td>
</tr>
</tbody>
</table>

Weighted Avg. 0,770 0,220 0,745 0,770 0,757 0,553 0,868 0,832

--- Confusion Matrix ---
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>213</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>54</td>
<td>280</td>
<td></td>
</tr>
</tbody>
</table>

--- classified as
<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>
Algorithm’s classification success rate in the case for our data is 77.0313%.

B. ANFIS application in MATLAB

For modelling membership functions, Parallel Coordinates Data Visualizer was used. Data set is not changed for this application.

![Parallel Coordinates data visualizer's figure of data set.](image)

Figure 5. Parallel Coordinates data visualizer’s figure of data set.

All input variables and output variable is used for modelling the wine quality. Figure 6 shows the fixed acidity’s membership functions. For all 11 input variables Bell membership function is used because of its being more successful than others.

![Fixed acidity membership functions.](image)

Figure 6. Fixed acidity membership functions.

Figure 7 shows the first 27 rules. In total 495 rule was used.
Figure 7. Fuzzy inference system’s first 27 rules.

Before training, model’s average testing error is 0.76674. Average testing error calculated as follows: Firstly model’s calculation for outputs are subtracted from data set’s outputs, then absolute values of these numbers are divided by 1599 (the total numbers of data in the set (Fig.8)).

After training, model’s average testing error is 0.46091 (Fig. 9).

To clarify inference system’s success, error tolerance should be specified. As Cortez’s paper [9] 0.5 and 1 is chosen for error tolerance for this paper too. With 1
error tolerance, 1532 of 1599 data (95.80%) are correctly classified; with 0.5 error tolerance, 1178 data (73.67%) are correctly classified. Here, 1 error tolerance means that if the absolute value of the difference between model’s output and data set’s output is less than or equal to 1, classification is accepted as true. 0.5 error tolerance means that if the absolute value of the difference between model’s output and data set’s output is less than or equal to 0.5, the classification is accepted as true.

4. CONCLUSION

This paper suggests two techniques for classification of wine quality. First one is random forest algorithm, with this algorithm 77.0313% classification success rate is achieved. It should not be forgotten that for this application data set is changed. The other one is ANFIS application. ANFIS is a system that have both fuzzy logic and artificial neural networks advantages. This application seems very satisfying when comparing with Cortez’s SVM (Support Vector Machine) regression model [9]. With 0.5 and 1 error tolerance SVM regression model accuracy is 62.4% and 89% respectively [9]. In this study, with 0.5 and 1 error tolerance, we obtain the accuracy as 73.67% and 95.80% respectively.

REFERENCES


Current address: A.E. OZALP: Hacettepe University, Faculty of Sciences, Dept. of Mathematics, Ankara, TURKEY
E-mail address: emre@gmail.com
ORCID Address: http://orcid.org/0000-0002-9170-3013

Current address: I.N. ASKERZADE: Ankara University, Faculty of Engineering, Dept. of Computer Engineering, Ankara, TURKEY
E-mail address: Iman.Askerbeyli@eng.ankara.edu.tr
ORCID Address: http://orcid.org/0000-0003-4466-8128