EXPLORING DATA MINING CLASSIFICATION APPROACH IN WEKA OPEN SOURCE

T. Chithrakumar[1] , Dr. M. Thangamani[2] & C. Premalatha[3]
[1]Assistant Professor, Department of IT, Sri Ramakrishna Engineering College, Coimbatore, India
[2]Assistant Professor, Kongu Engineering College, Perundurai, India
[3]Assistant Professor, Department of IT, Sri Ramakrishna Engineering College, Coimbatore, India

Abstract: The extraction of information from huge amount of data set is called data mining. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification predicts categorical labels. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka means Waikato Environment for Knowledge Analysis (WEKA). It is introduced by university of New Zealand and it has capacity to convert comma separated values file to relational table format. This research implementing the classification concept using Weka opens source using for WINE-QUALITY dataset.

Keywords – Data mining, Weka, Wine-Quality

1. INTRODUCTION

The data mining represents mining the knowledge from large data. Topics such as knowledge discovery, query language, decision tree induction, classification and prediction, cluster analysis, and how to mine the Web are functions of data mining. Manual analyses are time consuming in the real world. In this situation, WEKA can use for automating the task.

Weka is a collection of machine learning algorithms for data mining tasks. Classification was performed using WEKA in data mining research. WEKA is a data mining workbench that allows comparison between many different machine learning algorithms. In addition, it also has functionality for feature selection, data pre-processing and data visualization [1]. The algorithms can either be applied directly to a dataset or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Well-suited for developing new machine learning schemes. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

2. RELATED WORK

Various data mining classification concepts are discussed in [2-7]. WEKA enjoys widespread acceptance in both academia and business, has an active community, and has been downloaded more than 1.4 million times since being placed on Source Forge in April 2000. The customer datasets and bridge datasets are analyses using WEKA by [8,9]. Eibe Frank [10,11] highlighted a WEKA workbench and reviews the history of the project. Reena Thakur [12] presented data mining technology WEKA tool for the preprocessing, classification and analysis of in this institutional result of Computer science and engineering UG students.

3. EXPERIMENTS DESIGN

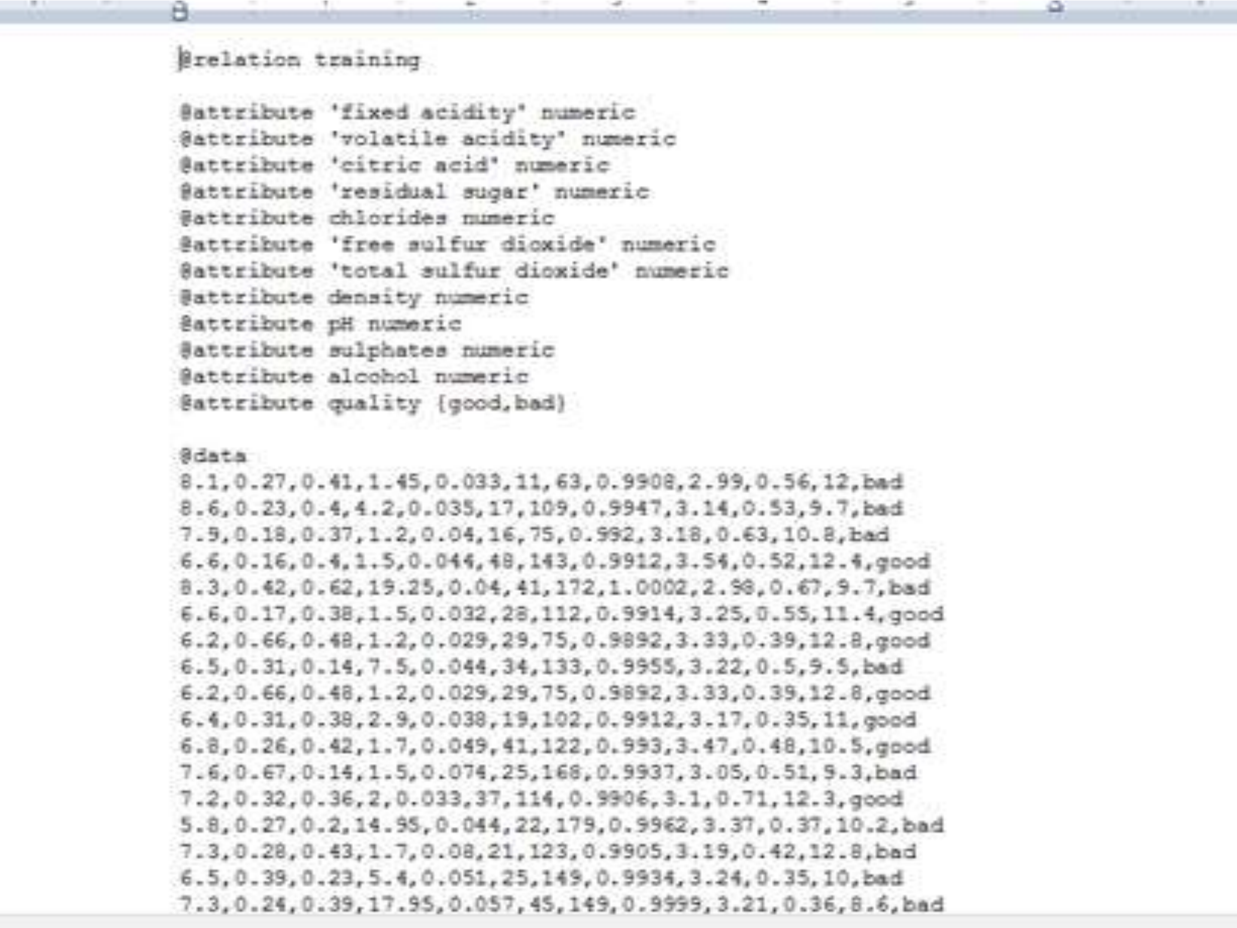
Use WEKA to build the classifier using WINE-QUALITY dataset by applying the classification algorithms (Nearest Neighbour classifier) and compare the results of the classifiers.

3.1 Dataset description

Data set Characteristics: Multivariate
Number of Instances: 1890
Number of Attributes: 12

3.2 Attributes description

The list of attributes in wine dataset are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol quality. Wine quality datasets are viewed as attribute file format and is illustrated in Fig. 1.

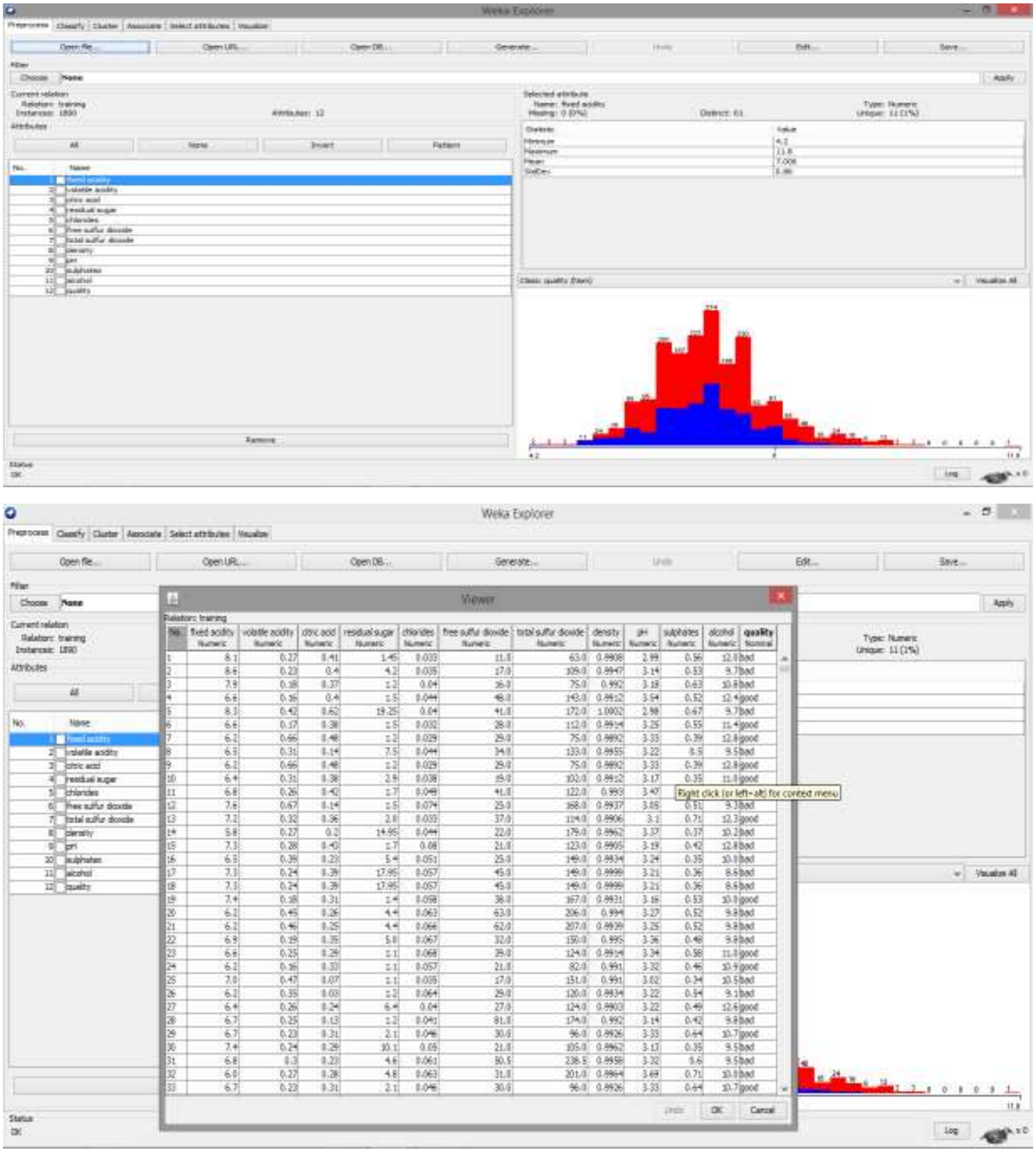


```
@relation training

@attribute 'fixed acidity' numeric
@attribute 'volatile acidity' numeric
@attribute 'citric acid' numeric
@attribute 'residual sugar' numeric
@attribute chlorides numeric
@attribute 'free sulfur dioxide' numeric
@attribute 'total sulfur dioxide' numeric
@attribute density numeric
@attribute pH numeric
@attribute sulphates numeric
@attribute alcohol numeric
@attribute quality {good,bad}

@data
8.1,0.27,0.41,1.45,0.033,11,63,0.9908,2.99,0.56,12,bad
8.6,0.23,0.4,4.2,0.035,17,109,0.9947,3.14,0.53,9.7,bad
7.9,0.18,0.37,1.2,0.04,16,75,0.992,3.18,0.63,10.8,bad
6.6,0.16,0.4,1.5,0.044,48,143,0.9912,3.54,0.52,12.4,good
8.3,0.42,0.62,19.25,0.04,41,172,1.0002,2.98,0.67,9.7,bad
6.6,0.17,0.38,1.5,0.032,28,112,0.9914,3.25,0.55,11.4,good
6.2,0.66,0.48,1.2,0.029,29,75,0.9892,3.33,0.39,12.8,good
6.5,0.31,0.14,7.5,0.044,34,133,0.9955,3.22,0.5,9.5,bad
6.2,0.66,0.48,1.2,0.029,29,75,0.9892,3.33,0.39,12.8,good
6.4,0.31,0.38,2.9,0.038,19,102,0.9912,3.17,0.35,11,good
6.8,0.26,0.42,1.7,0.049,41,122,0.993,3.47,0.48,10.5,good
7.6,0.67,0.14,1.5,0.074,25,168,0.9937,3.05,0.51,9.3,bad
7.2,0.32,0.36,2,0.033,37,114,0.9906,3.1,0.71,12.3,good
5.8,0.27,0.2,14.95,0.044,22,179,0.9962,3.37,0.37,10.2,bad
7.3,0.28,0.43,1.7,0.08,21,123,0.9905,3.19,0.42,12.8,bad
6.5,0.39,0.23,5.4,0.051,25,149,0.9934,3.24,0.35,10,bad
7.3,0.24,0.39,17.95,0.057,45,149,0.9999,3.21,0.36,8.6,bad
```

Fig.1 Wine quality datasets in attribute file format

4. IMPLEMENTATION STEPS

Many classification algorithms are available. The preprocessing WEKA is shown in Fig.2. In this stage remove the attribute id, since it uniquely identifies the tuples. It is done by selecting the remove attribute filter. Remove the attribute location, since it does not play a vital role in generating the rules. The Fig.3 represents the classification explorer panel in WEKA.



Filter: None

Current relation
Relation: training
Instances: 1890
Attributes: 12

Selected attribute
Name: fixed acidity
Missing: 0 (0%)
Distinct: 61
Type: Numeric
Unique: 11 (1%)

Statistic	Value
Minimum	4.2
Maximum	11.8
Mean	7.008
StdDev	0.86

No.	Name
1	fixed acidity
2	volatile acidity
3	citric acid
4	residual sugar
5	chlorides
6	free sulfur dioxide
7	total sulfur dioxide
8	density
9	pH
10	sulphates
11	alcohol
12	quality

Class: quality (Nom)

Status
OK

Viewer

Relation: training

No.	fixed acidity Numeric	volatile acidity Numeric	citric acid Numeric	residual sugar Numeric	chlorides Numeric	free sulfur dioxide Numeric	total sulfur dioxide Numeric	density Numeric	pH Numeric	sulphates Numeric	alcohol Numeric	quality Nominal
1	8.1	0.27	0.41	1.45	0.033	11.0	63.0	0.9908	2.99	0.56	12.0	bad
2	8.6	0.23	0.4	4.2	0.035	17.0	109.0	0.9947	3.14	0.53	9.7	bad
3	7.9	0.18	0.37	1.2	0.04	16.0	75.0	0.992	3.18	0.63	10.8	bad
4	6.6	0.16	0.4	1.5	0.044	48.0	143.0	0.9912	3.54	0.52	12.4	good
5	8.3	0.42	0.62	19.25	0.04	41.0	172.0	1.0002	2.98	0.67	9.7	bad
6	6.6	0.17	0.38	1.5	0.032	28.0	112.0	0.9914	3.25	0.55	11.4	good
7	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	12.8	good
8	6.5	0.31	0.14	7.5	0.044	34.0	133.0	0.9955	3.22	0.5	9.5	bad
9	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	12.8	good
10	6.4	0.31	0.38	2.9	0.038	19.0	102.0	0.9912	3.17	0.35	11.0	good
11	6.8	0.26	0.42	1.7	0.049	41.0	122.0	0.993	3.47	[illegible]	[illegible]	[illegible]
12	7.6	0.67	0.14	1.5	0.074	25.0	168.0	0.9937	3.05	0.51	9.3	bad
13	7.2	0.32	0.36	2.0	0.033	37.0	114.0	0.9906	3.1	0.71	12.3	good
14	5.8	0.27	0.2	14.95	0.044	22.0	179.0	0.9962	3.37	0.37	10.2	bad
15	7.3	0.28	0.43	1.7	0.08	21.0	123.0	0.9905	3.19	0.42	12.8	bad
16	6.5	0.39	0.23	5.4	0.051	25.0	149.0	0.9934	3.24	0.35	10.0	bad
17	7.3	0.24	0.39	17.95	0.057	45.0	149.0	0.9999	3.21	0.36	8.6	bad
18	7.3	0.24	0.39	17.95	0.057	45.0	149.0	0.9999	3.21	0.36	8.6	bad
19	7.4	0.18	0.31	1.4	0.058	38.0	167.0	0.9931	3.16	0.53	10.0	good
20	6.2	0.45	0.26	4.4	0.063	63.0	206.0	0.994	3.27	0.52	9.8	bad
21	6.2	0.46	0.25	4.4	0.066	62.0	207.0	0.9939	3.25	0.52	9.8	bad
22	6.9	0.19	0.35	5.0	0.067	32.0	150.0	0.995	3.36	0.48	9.8	bad
23	6.6	0.25	0.29	1.1	0.068	39.0	124.0	0.9914	3.34	0.58	11.0	good
24	6.2	0.16	0.33	1.1	0.057	21.0	82.0	0.991	3.32	0.46	10.9	good
25	7.0	0.47	0.07	1.1	0.035	17.0	151.0	0.991	3.02	0.34	10.5	bad
26	6.2	0.35	0.03	1.2	0.064	29.0	120.0	0.9934	3.22	0.54	9.1	bad
27	6.4	0.26	0.24	6.4	0.04	27.0	124.0	0.9903	3.22	0.49	12.6	good
28	6.7	0.25	0.13	1.2	0.041	81.0	174.0	0.992	3.14	0.42	9.8	bad
29	6.7	0.23	0.31	2.1	0.046	30.0	96.0	0.9926	3.33	0.64	10.7	good
30	7.4	0.24	0.29	10.1	0.05	21.0	105.0	0.9962	3.13	0.35	9.5	bad
31	6.8	0.3	0.23	4.6	0.061	50.5	238.5	0.9958	3.32	0.6	9.5	bad
32	6.0	0.27	0.28	4.8	0.063	31.0	201.0	0.9964	3.69	0.71	10.0	bad
33	6.7	0.23	0.31	2.1	0.046	30.0	96.0	0.9926	3.33	0.64	10.7	good

Right click (or left-alt) for context menu

Undo | OK | Cancel

Fig.2 preprocessing the wine quality data set in WEKA Explorer Panel

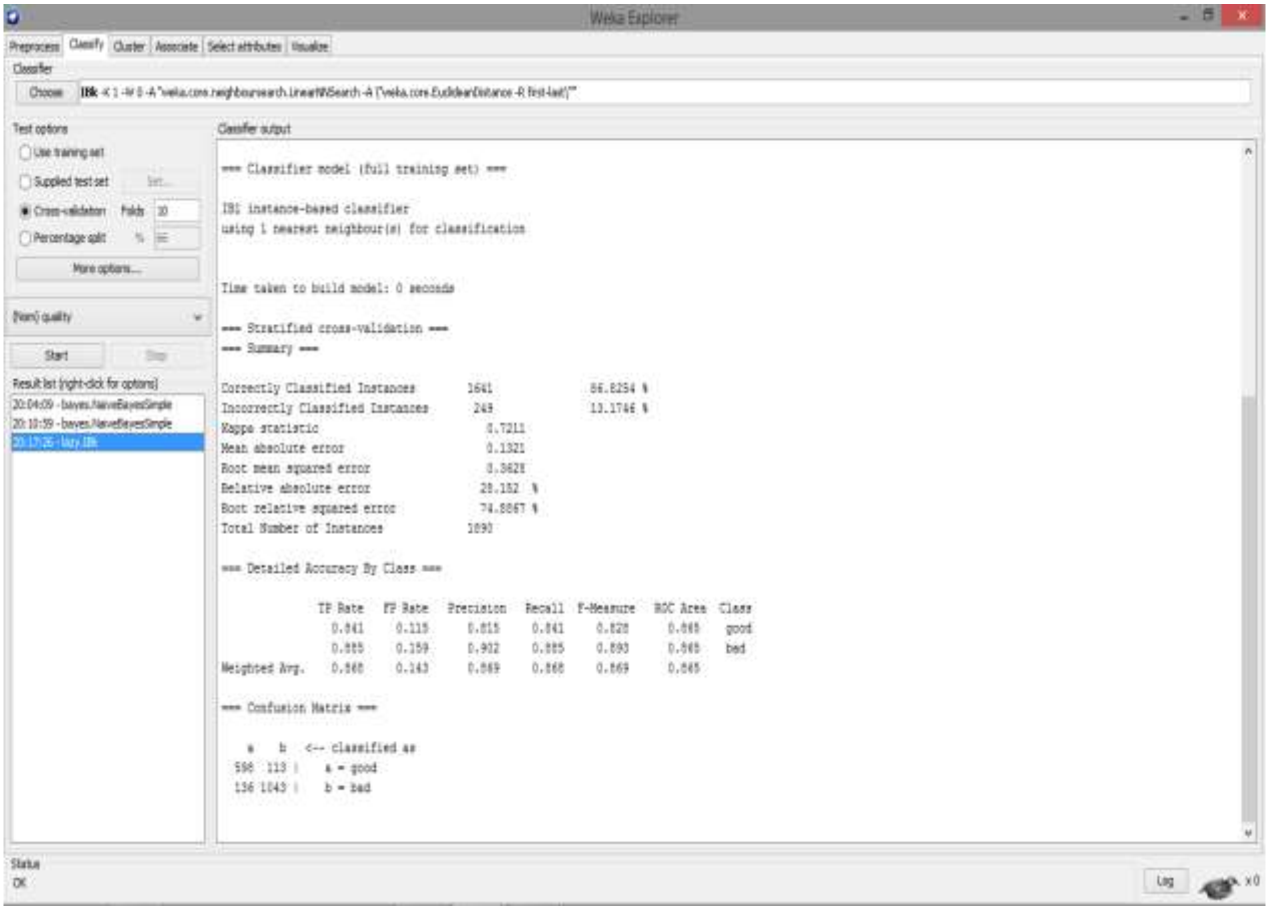
10 fold Cross Validation test is applied to k-Nearest Neighbour classifier for wine quality datasets. Ten fold cross validation means data set is divided into 10 equal parts. Nine fold is used for training and remaining one fold for testing.

5. EXPERMENT RESULT

WEKA classifier starts to learn by clicking start button in classifier panel. After learning, it builds classifier and produce result in classifier output panel. It shows what type of relation used, how many attributes in the relation and also displays list of attributes. It shows what type of test used for what type of algorithms.

Confusion matrix produce correctly classified instance and incorrectly classified instance in the matrix format. Diagonal elements are treated as correctly classified instance and remaining are incorrectly classified instance. Fig.4 shows result of the k-Nearest Neighbour classifier for wine quality data sets.

Fig.3 k-Nearest Neighbour classifier in WEKA using 10 fold cross validation



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) quality

Result list (right-click for options)
20:04:09 - bayes.NaiveBayesSimple
20:10:59 - bayes.NaiveBayesSimple
20:17:26 - lazy.IBk

```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1641               86.8254 %
Incorrectly Classified Instances       249               13.1746 %
Kappa statistic                          0.7211
Mean absolute error                      0.1321
Root mean squared error                  0.3628
Relative absolute error                 28.152  %
Root relative squared error             74.8867 %
Total Number of Instances             1890

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.841     0.115      0.815     0.841     0.828      0.865    good
                 0.885     0.159      0.902     0.885     0.893      0.865    bad
Weighted Avg.    0.868     0.143      0.869     0.868     0.869      0.865

=== Confusion Matrix ===

    a    b   <-- classified as
  598  113 |    a = good
  136 1043 |    b = bad
```

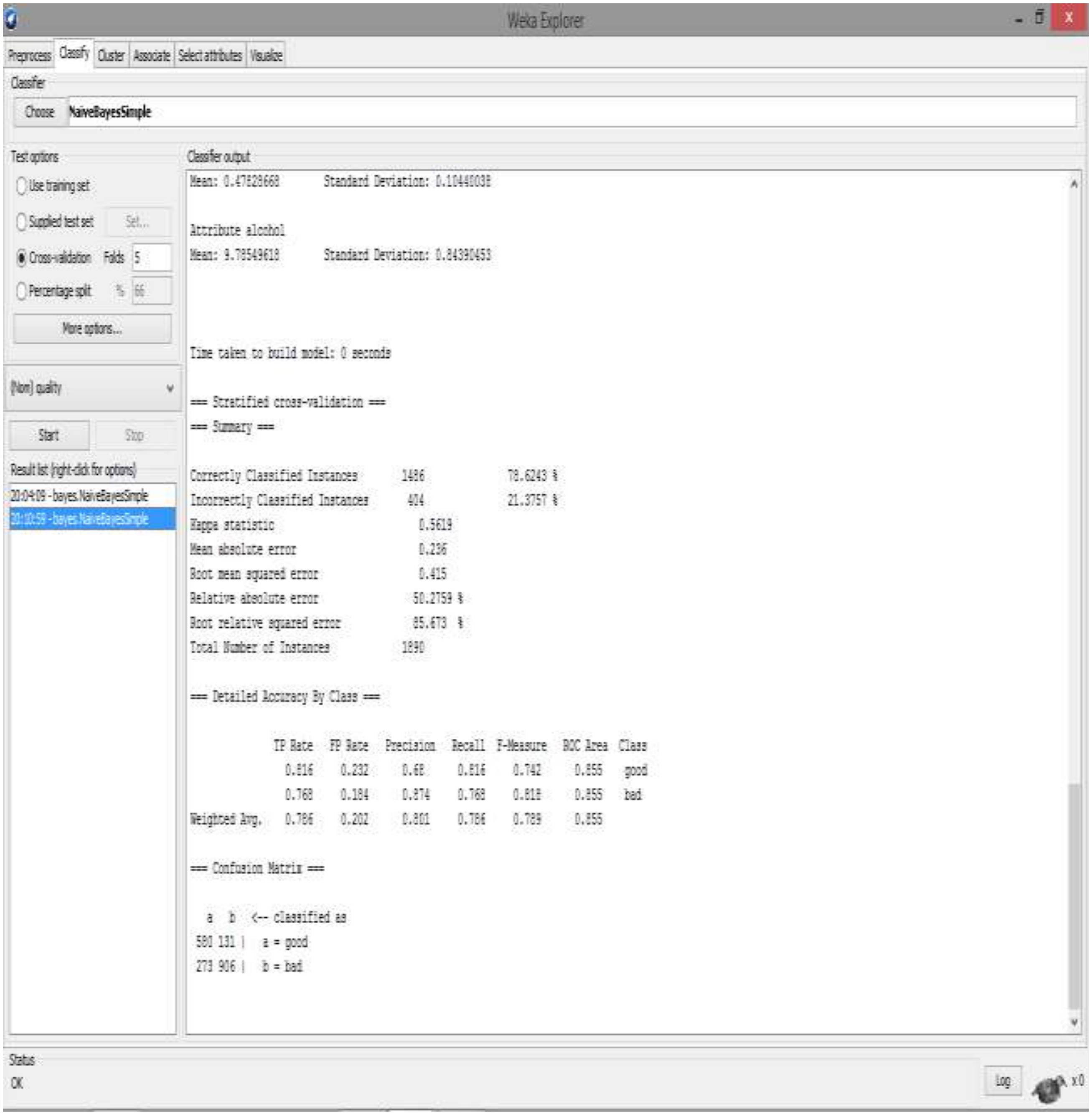
Status
OK

=== Run information ===

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -
A \"weka.core.EuclideanDistance -R first-last\""
Relation:     training
Instances:    1890
Attributes:   12
              fixed acidity
              volatile acidity
              citric acid
              residual sugar
              chlorides
              free sulfur dioxide
              total sulfur dioxide
              density
              pH
              sulphates
              alcohol
              quality
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1641               86.8254 %
Incorrectly Classified Instances       249               13.1746 %
Kappa statistic                          0.7211
Mean absolute error                      0.1321
Root mean squared error                  0.3628
Relative absolute error                 28.152  %
Root relative squared error             74.8867 %
Total Number of Instances             1890
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.841     0.115     0.815      0.841    0.828        0.865    good
                 0.885     0.159     0.902      0.885    0.893        0.865    bad
Weighted Avg.    0.868     0.143     0.869      0.868    0.869        0.865

=== Confusion Matrix ===

   a     b   <-- classified as
 598   113 |   a = good
```

```
  136    1043 |    b = bad
```

5- fold Cross Validation test is applied to k-Nearest Neighbour classifier using Wine quality datasets is shown in Fig.4 and classifier evaluation is illustrated in Fig.5.



Classifier: NaiveBayesSimple

Test options:
○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 5
○ Percentage split   % 66

(Nom) quality

Result list (right-click for options)
20:04:09 - bayes.NaiveBayesSimple
20:10:59 - bayes.NaiveBayesSimple

```
Mean: 0.47828668       Standard Deviation: 0.10448038


Attribute alcohol
Mean: 9.78549618       Standard Deviation: 0.84390453




Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1486              78.6243 %
Incorrectly Classified Instances       404              21.3757 %
Kappa statistic                          0.5619
Mean absolute error                      0.236
Root mean squared error                  0.415
Relative absolute error                 50.2759 %
Root relative squared error             85.673  %
Total Number of Instances             1890

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.816     0.232     0.68      0.816     0.742      0.855    good
                 0.768     0.184     0.874     0.768     0.818      0.855    bad
Weighted Avg.    0.786     0.202     0.801     0.786     0.789      0.855

=== Confusion Matrix ===

   a   b   <-- classified as
 580 131 |   a = good
 273 906 |   b = bad
```

Fig.4 Building classifier using 5 fold cross validation

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -
A \"weka.core.EuclideanDistance -R first-last\""
Relation:     training
Instances:    1890
Attributes:   12
              fixed acidity
              volatile acidity
              citric acid
              residual sugar
              chlorides
              free sulfur dioxide
              total sulfur dioxide
              density
              pH
              sulphates
              alcohol
              quality
Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1631               86.2963 %
Incorrectly Classified Instances       259               13.7037 %
Kappa statistic                          0.7091
Mean absolute error                      0.1375
Root mean squared error                  0.37
Relative absolute error                 29.2889 %
Root relative squared error             76.3703 %
Total Number of Instances               1890

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.827     0.115      0.812     0.827     0.82        0.858    good
                   0.885     0.173      0.895     0.885     0.89        0.858    bad
```

```
Weighted Avg.    0.863    0.151    0.864     0.863   0.863        0.858

=== Confusion Matrix ===
   a       b   <-- classified as
 588     123 |    a = good
 136    1043 |    b = bad
```

Table 1 Evaluation of classifier

Classifier	Time taken to build model	Test mode	Correctly classified instances	Incorrectly classified instances	Kappa Statistic	Mean absolute error	Root Mean squared error	Relative absolute error	Root relative squared error
Lazy-IBk	0 Seconds	10-Fold Cross Validation	1641/1890 (86.83%)	249/1890 (13.17%)	0.7211	0.1321	0.3628	28.15%	74.89%
Lazy-IBk	0 Seconds	5-Fold Cross Validation	1631/1890 (86.30%)	259/1890 (13.70%)	0.7091	0.1375	0.37	29.29%	76.37%

CONCLUSION:

In this paper provides information about how raw data can be transformed into meaningful information. Data sets are tested with different cross validation. In future, it can build the various classifiers and compare the classifiers with same and different data sets.

1. Donn Morrison, Ruili Wang, Liyanage C. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech Communication, Elsevier, Vol. 49, pp.98-112, 2007
2. Han, J., Kamber, M., Jian P., Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers, 2011.
3. Giraud Carrier,C., and Povel, O., Characterising Data Mining software, Intelligent Data Analysis, Vol.7 No.3,Pp.181-192, 2003
4. P. Brazdil, C. Soares, and J. Da Costa. Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results, Machine Learning, Vol.50, No.3, Pp.251–277, 2003.
5. L. Xu, H. H. Hoos, and K. Leyton-Brown. Hydra:
6. Automatically configuring algorithms for portfolio-based selection. In Proc. of AAAI, Vol 10, Pp.210-216.
7. Guerra L, McGarry M, Robles V, Bielza C, Larrañaga P, Yuste R. Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. Developmental neurobiology, Vol.7, No.1, Pp. 71-82, 2011.
8. Dr. M. Thangamani & V.Prasanna. Implementation of Association Rule Mining for Bridge Datasets Using Weka, International Research Journal in Global Engineering and Sciences, Vol.1. Issue 1. pp. 1-13, 2016

9. N.Suresh Kumar, Dr. M. Thangamani. Effective Customer Patterns Analysis Using Open Source Weka Data Mining Tool, International Research Journal in Global Engineering and Sciences (IRJGES), 2016, Vol.1. Issue 1. pp. 14-33
10. Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Vol.11, No.1, Pp.1-18, 2010.
11. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. Journal of Machine Learning. Research, Vol.9 Pp.1871–1874, 2008.
12. Reena Thakur. A.R.Mahajan, Preprocessing and Classification of Data Analysis in Institutional System using Weka, International Journal of Computer Applications, Vol.112, No. 6, Pp. 9-11, 2015.

Authors Biography

Mr.T.Chithrakumar is currently working as Assistant Professor in Department of Information Technology. He obtained his M.E. degree in Computer Science and Engineering from V.S.B College of Engineering Technical Campus, Coimbatore in the year 2015 and completed his B.Tech. Degree in Information Technology from Ranganathan Engineering College, Coimbatore in the year 2011. His areas of interest include Data Mining, Networks, Cloud Computing, Mobile Adhoc Network. He has Teaching Experience of 9 Months at Sri Ramakrishna Engineering College, Coimbatore. He has published papers in the area of Networking, Cloud Computing and Adhoc networks in National/International conferences and reputed journals.

Dr. M. Thangamani possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 70 articles in refereed and indexed journals, books and book chapters and presented over 67 papers in national and international conferences in above field. She has delivered more than 79 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got best paper awards from various education related social activities in India and Abroad. She is on the editorial board and reviewing committee of leading research journals, which includes her nomination as the Editor in chief to International Scientific Global journal for Engineering, Science and Applied Research (ISGJESAR) & International Research Journal in Global Engineering and Sciences (IRJGES) and on the program committee of top international data mining and soft computing conferences in various countries.

Ms. C. Premalatha is currently working as Assistant Professor in the Department of Information Technology at Sri Ramakrishna Engineering College, Coimbatore. She has a master’s degree in Computer Science and Engineering (2014) from Ranganathan Engineering College and bachelor’s degree in Information Technology (2012) from Sri Ramakrishna Engineering College. Her teaching experience spans over 1 year and 7 months. Her research and teaching interests include Green computing, Knowledge Discovery and Mining (Classification Techniques) and Security over internet and intranet. She has published papers in data mining in National Conferences and International Journals. She is a life member of IAENG, IACSIT, CSTA and ICST.