ISSN: 0974-6471, Vol. 10, No. (4) 2017, Pg. 788-792

Oriental Journal of Computer Science and Technology

Journal Website: www.computerscijournal.org

A Comparative Analysis of Classification Algorithms on Weather Dataset Using Data Mining Tool

D. RAMESH*, SYED NAWAZ PASHA and G. ROOPA

Department of CSE, S R Engineering College Warangal 506371, Telangana, India.

Abstract

Data mining has become one of the emerging fields in research because of its vast contents. Data mining is used for finding hidden patterns in the database or any other information repository. This information is necessary to generate knowledge from the patterns. The main task is to extract knowledge out of the information. In this paper we use a data mining technique called classification to determine the playing condition based on the current temperature values. Classification technique is a powerful way to classify the attributes of the dataset into different classes. In our approach we use classification algorithms like Decision Tree (J48), REP Tree and Random Tree. Then we compare the efficiencies of these classification algorithms. The tool we use for this approach is WEKA (Waikato Environment for Knowledge Analysis) a collection of open source machine learning algorithms.

Article History

Received: 17 October 2017
Accepted:11 December 2017

Keywords

Data Mining,
Data set,
Classification.

Introduction

Data mining is that the method to extract or mine data from immense volume of information. Broadly data processing will be outlined because the task of extracting implicit, antecedently unknown potential helpful data from knowledge in giant databases. Data mining tasks are classified as descriptive which discover interesting patterns or relationships describing the data and predictive task which predicts or classifies the behavior of the model supported obtainable information. It's a content field with a general goal of predicting outcomes and uncovering relationships. Some of the data mining techniques are Classification, Clustering and Rule Mining.

Clustering is that the most typically used information discovery technique. It helps un-covering the unknown category labels. It helps un-covering the unknown class labels. Clustering has gained importance in many applications in the recent past. Most of the cluster algorithms area unit ascendable to large dataset. Weather is random entity. Forecasting is the technology to predict the atmosphere at given location and a given time taking into consideration various factors such as humidity, temperature, wind

CONTACT RAMESH DADI dadiramesh44@gmail.com Department of CSE, S R Engineering College Warangal 506371,Telangana, India.

To link to this article: http://dx.doi.org/10.13005/ojcst/10.04.13

and outlook. It's done by gathering the information regarding this state of the atmosphere at a given location thus applies scientific understanding to predict but the temperature will modification over the course of some time. In our paper we are going to predict whether the play can happen based on current weather values such as temperature, humidity, windy, outlook[11]. We make the prediction based on various classification algorithms such as Decision Tree (J48), REP Tree and Random Tree. We conjointly compare every of those algorithms in terms of their accuracy mistreatment completely different measures.

Classification Algorithms

Decision Tree Induction

DTI is a tree learning algorithms. It consists of flow diagram like structure wherever the inner node denotes a take a look at on the attribute, the branches will denote the outcome of the test performed on the attribute and the leaf nodes will denote class labels.

The internal nodes are represented as rectangles and the leaf nodes are represented with oval shapes.

To determine the cacophonic attribute it makes use of various attribute choice measures like data gain, gain quantitative relation and Gini Index.

Example: J48,C 4.5,CART

REP Tree

It is a decision tree learner algorithm. It constructs the decision tree exploitation data gain or variance then prunes it exploitation reduced error pruning exploitation back fitting strategy. REP Tree Iteratively generates multiple trees using regression logic. It sorts the values for numeric attribute only once. It deals with missing values by rending the corresponding instances into items.

Random Trees

This algorithm can deal with both regression and classification problems. it's a group of tree predictors that's referred to as forest. It takes the input as feature vector and compares it with each tree within the forest and offers the result category label that has highest votes.

Classifier Output Measures

The classifier classifies the tuples in the dataset. It is quite natural that the classifier may have error rate and may fail to correctly classify the tuples.Hence we measure the classifier accuracy which is given by the percentage of instances that square measure properly classified by classifier.

Confusion Matrix

It gives information regarding the classifier output in terms of the number of tuples that are correctly classified and the numbers of tuples that are miss classified. For a good accuracy classifier the elements must be in along the diagonal while the other entries must be close to zero.

Mean Absolute Error

It is a measure for accuracy. It is the mean of the absolute errors that is the mean of the distinction between the expected value and also the actual value.

Root Mean Square Error

If we take the square root of the mean square error then we obtain the root mean square error. We do it to adjust large error rates.

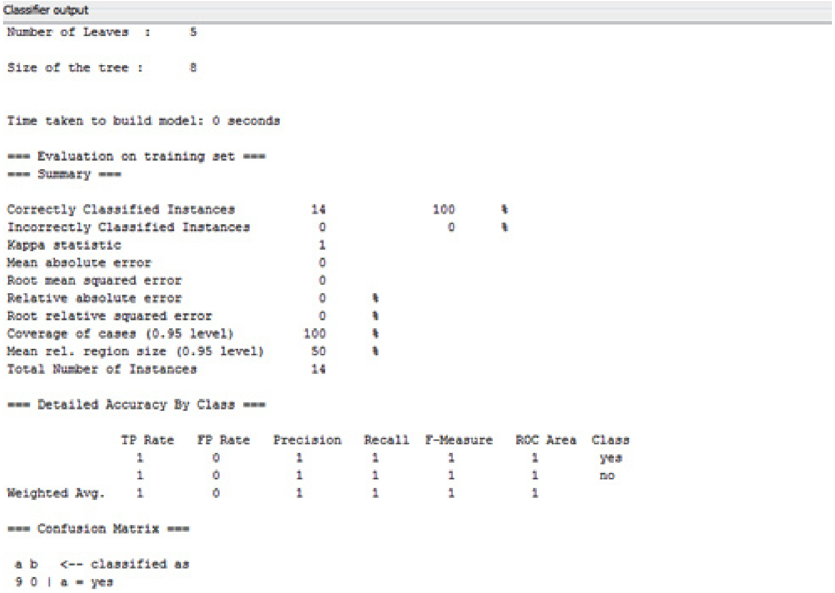
Results and Comparisons

The tool we used for the result analysis is WEKA which consists of large number of open source machine learning algorithms. It takes the input in the form of ARFF (Attribute Relation File Format),CSV(comma separated values).The data set we used is weather which is input to weka in ARFF format.

The weather data set contains following attributes.

Wind	{yes,no}
Temperature	{hot,cool,mild}
Humidity	{high,normal}
Outlook	{sunny,overcast,rainy}
Play	{yes,No}

Result of J48 classifier

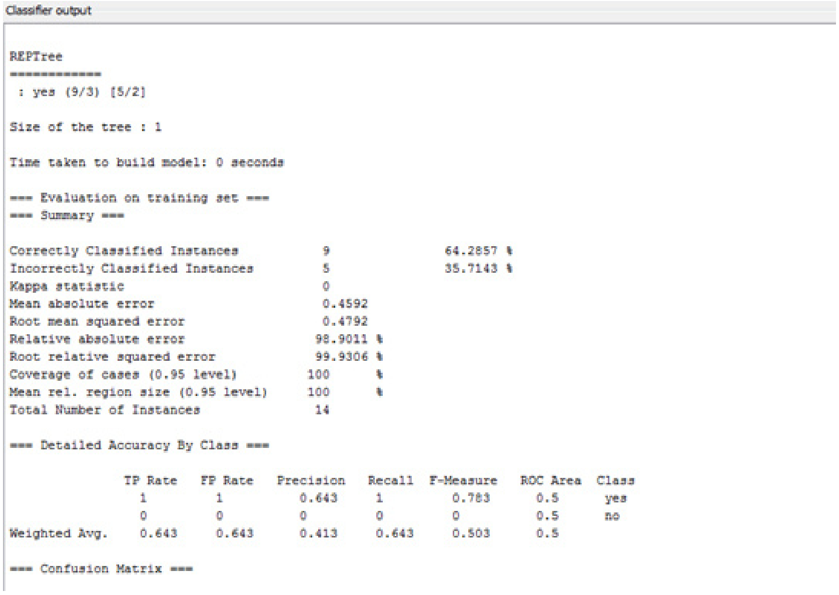


Classifier output

```
Number of Leaves  :     5

Size of the tree :      8


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          14              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        yes
                 1         0          1         1         1          1        no
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
```

Fig. 1: Statistics of J48 classifier on weather dataset

Result of REP Tree Classifier

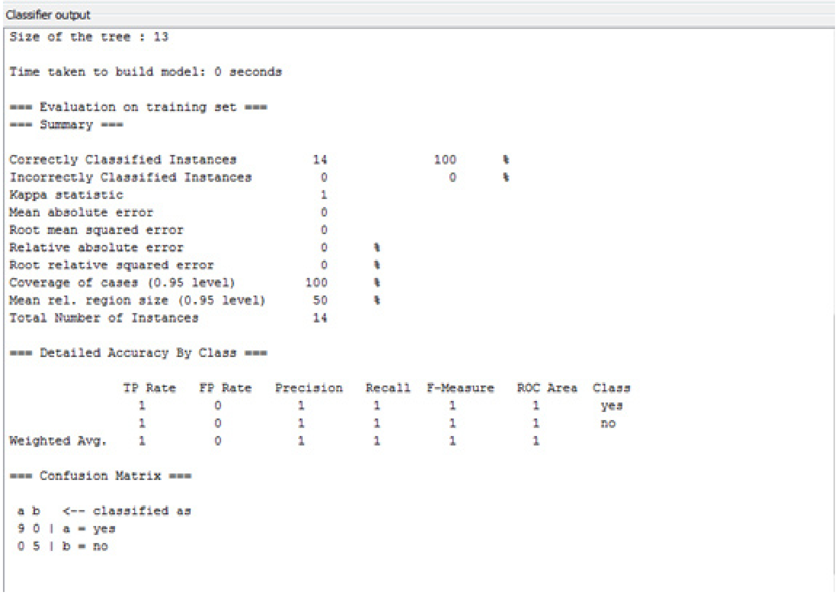


Classifier output

```
REPTree
============
 : yes (9/3) [5/2]

Size of the tree : 1

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0
Mean absolute error                      0.4592
Root mean squared error                  0.4792
Relative absolute error                 98.9011 %
Root relative squared error             99.9306 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.643     1         0.783      0.5      yes
                 0         0          0         0         0          0.5      no
Weighted Avg.    0.643     0.643      0.413     0.643     0.503      0.5

=== Confusion Matrix ===
```

Fig. 2: Statistics of REP Tree classifier on weather dataset

Result of Random Tree



```
Size of the tree : 13

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          14              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0       %
Root relative squared error              0       %
Coverage of cases (0.95 level)         100       %
Mean rel. region size (0.95 level)      50       %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0         1          1         1          1        yes
                 1         0         1          1         1          1        no
Weighted Avg.    1         0         1          1         1          1

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

Fig. 3:Statistics of Random Tree classifier on weather dataset

Table 1: Overall comparison of J48, REP Tree and Random Tree (using training dataset)

Classifier/ Results	Total number of Instances	Correctly classified Instances	Miss classified Instances	accuracy	Mean Absolute Error	Root Mean Squared Error
J48	14	14	0	100%	0	0
REP Tree	14	9	5	64.28%	0.4592	0.4792
Random Tree	14	14	0	100%	0	0

Conclusion

This paper intends to study the classifier accuracy of various classification algorithms using WEKA tool on weather dataset.The experimental results of the various classification algorithms is listed.First the experiment was done on the weather dataset using j48 algorithm which classifies all the instances correctly.The accuracy of the j48 classifier is 100%. Then the dataset was run on Random Tree classifier which classifies all instances correctly and has 100 % accuracy.Then classification was done using REP Tree classifier and we found the accuracy was decreased to 64.28 % because it was not able to classify all the instances correctly and we found that 5 instances were misclassified by REP Tree classifier because of which its accuracy is decreased.

References

1 Germano C. Vasconcelos, Paulo J. L. Adeodato and Domingos S. M. P. Monteiro. 1999. A Neural Network Based Solution for the Credit Risk Assessment Problem.

Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais, (July 1999), 269-274.

2 Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen. 2002. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications (Elsevier)* **23**, 245–254.

3 Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA", International Conference in "*Emerging Trends in Science, Technology and Management*,2013

4 S.Archana1, Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", *International Journal of Computer Science and Mobile Applications,* Vol.**2** Issue. 2, February- 2014

5 Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu. 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study," *Decision Support Systems (Elsevier)* **37**, 543– 558.

6 Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. 2006. Credit Risk Analysis Using a ReliabilityBased Neural Network Ensemble Model. S. Kollias et al. (Eds.): ICANN 2006, Part II, Springer LNCS 4132, 682 – 690.

7 Eliana Angelini, Giacomo di Tollo, and Andrea Roli. 2006. A Neural Network Approach for Credit Risk Evaluation," Kluwer Academic Publishers, 1 – 22.

8 S. Kotsiantis. 2007. Credit risk analysis using a hybrid data mining model. *Int. J. Intelligent Systems Technologies and Applications,* Vol. **2**, No. 4, 345 – 356.

9 Hamadi Matoussi and Aida Krichene. 2007. Credit risk assessment using Multilayer Neural Network Models - Case of a Tunisian bank.

10 Lean Yu, Shouyang Wang, and Kin Keung Lai. 2008. Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications (Elsevier)* **34**, pp.1434–1444.

11 Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi. 2011. A Hybrid Method for Credit Risk Assessment of Bank Customers. *International Journal of Trade, Economics and Finance*, Vol. **2**, No. 2, (April 2011)

12 UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.

13 Tina R. Patil, and S. S. Sherekar. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications* Vol. **6**, No.2, (Apr 2013), 256 - 261.

14 Witten IH, and Frank E. 2005. Data mining: practical machine learning tools and techniques – 2nd ed. the United States of America, Morgan Kaufmann series in data management systems.

15 Quinlan J (1987) Simplifying decision trees, *International Journal of Man Machine Studies,* **27**(3), 221–234.

16 S.K. Jayanthi and S.Sasikala. 2013. REPTree Classifier for indentifying Link Spam in Web *Search Engines. IJSC,* Volume **3**, Issue 2, (Jan 2013), 498 – 505.

17 Leo Breiman. 2001. Random Forests. *Machine Learning*. **45**(1): 5-32.

18 Margaret H. Danham, and S. Sridhar. 2006. Data mining, Introductory and Advanced Topics. Person education, 1st Edition

19 Lakshmi Devasena, C. 2014. Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction. *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. **3**, Issue 4, 6156 – 6162

20 Bhavani M, Vinod Kumar S "A data mining approach for precise diagnosis of dengue fever", *International journal of latest trends in engineering and technology*,vol. **7**, issue 4. 2016