Global Journal of Computer Science and Technology: H Information & Technology
Volume 20 Issue 2 Version 1.0 Year 2020
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Comparative Analysis of Random Forest and J48 Classifiers for "IRIS" Variety Prediction

By Youssef Fakir, Youness Lakhdoura & Rachid Elayachi
Sultan Moulay Slimane University

Abstract- Data mining may be a computerized technology that uses complicated algorithms to seek out relationships and trends in large databases, real or perceived, previously unknown to the retailer, to market decision support. Data mining is predicted to be one of the widespread recognition of the potential for analysis of past transaction data to enhance the standard of future business decisions. The aim is to arrange a set of knowledge items and classify them.

In this paper, we apply two classifier algorithms: J48 (c4.5) and Random Forest on the IRIS dataset, and we compare their performance based on different measures.

Keywords: IRIS, J48 classifier, proficiency comparison, random forest classifier.

GJCST-H Classification: J.1

Strictly as per the compliance and regulations of:

Comparative Analysis of Random Forest and J48 Classifiers for "IRIS" Variety Prediction

Youssef Fakir [α], Youness Lakhdoura [σ] & Rachid Elayachi [ρ]

Abstract- Data mining may be a computerized technology that uses complicated algorithms to seek out relationships and trends in large databases, real or perceived, previously unknown to the retailer, to market decision support. Data mining is predicted to be one of the widespread recognition of the potential for analysis of past transaction data to enhance the standard of future business decisions. The aim is to arrange a set of knowledge items and classify them.

In this paper, we apply two classifier algorithms: J48 (c4.5) and Random Forest on the IRIS dataset, and we compare their performance based on different measures.

Keywords: IRIS, J48 classifier, proficiency comparison, random forest classifier.

I. Introduction

People are often susceptible to making mistakes during analyses or, possibly, when trying to determine relationships between multiple features. This fact, makes it difficult for them to seek out solutions to certain problems. Data mining involves the utilization of sophisticated data analysis tools to get previously unknown, valid patterns, and relationships in the datasets[1]. These tools can include statistical models, mathematical algorithms, and machine learning methods [2].

Consequently, data processing consists of quite a collection and managing data, it also includes analysis and prediction [1].

The classification technique is capable of processing a sort of data than regression and is growing in popularity [3].

II. Dataset Used

In this research work, we use the IRIS plant data set, one of the most popular databases for the classification problems, it is obtained from UCI Machine Learning Repository and created by R.A. Fisher while donated by Michael Marshall (MARSHALL%PLU @io.arc.nasa.gov) on July 1988[4].

The IRIS dataset contains three different classes of IRIS plants depending on their pattern [5,6]. Each class of IRIS plant contain fifty objects. The attributes that already predicted belongs to a category of IRIS plant. The list of attributes presents within the IRIS is often described as categorical, nominal, and continuous. The experts have mentioned that the info set is complete i.e. there isn't any missing value found in any attribute of this data set [6].

This research makes use of the documented IRIS dataset, which contains three classes of fifty instances each. The 150 instances, which are equally divided between the three classes, hold the subsequent four numeric attributes:

1. Sepal length - continuous
2. Sepal width - continuous
3. Petal length - continuous
4. Petal width –continuous

And therefore the fifth attribute "Variety" is that the predictive attribute which identifies which class of the following belongs the instance: IRIS Setosa, IRISVersicolor, or IRIS Virginica [5,6].

III. Classifiers Used

In this paper, we compared the proficiency assessment of IRIS variety for two tree based classifiers: Random Forest and J48 Classifiers.

a) Random Forest Classifier

Random Forest [7] is considered one of the best "off-the-shelf" classifiers for high-dimensional data. Random forest is a mix of tree predictors sampled autonomously count on the values of a random vector following an equivalent distribution for all trees of the forest. The generalization error of random forest classifier depends on the association between the individual trees inside the forest and the strength of them. The dataset divided into a training dataset to learn each tree, and the remaining of the data set is used to estimate error and variable importance. Class assignment is formed according to the number of votes for any of the trees, to apply the model of the results. it's almost like bagged decision trees with hardly some key differences as given below:

For every split point, the search isn't overall p variables but just over m (number of tested) variables (where, e.g,m = [p/3])

No pruning necessary. Trees are often grown until each node contains just only a few observations.
The Random Forest gave better prediction, and almost no parameter adjustment is necessary.

Author α σ ρ: Computer science department, Science and Technique faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco.
e-mails: fakfad@yahoo.fr, lakhdouryouness@gmail.com, info.dec07@yahoo.fr

b) J48 Classifier

The J48 classifier is an extension of the decision tree C4.5 algorithm for classification [8], which creates a binary tree. It's the foremost useful decision tree approach for classification problems. This system constructs a tree to model the classification process. After the tree is made, the algorithm is applied to every tuple within the database and leads to classification for that tuple [9].

Algorithm J48 [9]:

```
INPUT:
P//Training data
OUTPUT
DT //Decision tree
DTBUILD (*P)
{
DT=φ;
DT= Create root node and label with splitting attribute;
DT= Add arc to root node for each split
predicate and label;
For each arc do
P= Database created by applying spliting
predicate to P;
If stopping point reached for this path, then
DT'= create leaf node and label with
appropriate class;
Else
DT'= DTBUILD(P);
DT= add DT' to arc;
}
```

The absent values are ignored byJ48 while building a decision tree, i.e. the known information about the attribute values for the other records is helpful to predict the value for that item. The idea is to divide the data into a range based on the attribute values for that element which are identified in the training sample [10].

IV. Performance Measures Used

Various scales are wont to gauge the performance of the classifiers.

a) Classification Accuracy (CA)

Classification accuracy presents the percent of correctly classified instance in the test dataset. We calculate it by dividing the correctly classified instances by the total number of instance multiplied by 100.

b) Mean Absolute Error (MAE)

Mean absolute error is that the average of the variance between predicted and actual value altogether test cases. It's an honest measure to measure performance.

c) Root Mean Square Error (RMSE)

Root mean squared error is employed to scale dissimilarities between values. It's determined by taking the root of the mean square error.

d) Confusion Matrix (CM)

A confusion matrix is a tool checking in particular how often the predictions are correct compared to reality in classification problems.

V. Results and Discussion

In this work, to evaluate the performance of the different Tree-based Classifiers (Random Forest and J48), we used a well-known open-source tool in the machine learning field called "WEKA". The performance is tested using two methods, first by splitting the dataset into training (70%) and testing (30%) datasets, as well as using different Cross-Validation methods.

a) Performance of Random Forest Classifier

Table 1 show the global evaluation summary of Random Forest Classifier using both of the test modes: splitting and different cross-validation methods. Fig.1 and Fig.2 display the performance of Random Forest Classifier in terms of Classification Accuracy and time taken to build the model. From Table I to Table VI we gave the confusion matrix for different test modes.

By applying these test modes using Random Forest Classifier, we got 95.55% accuracy, spending 0.17s on building the model for the split. Using different cross-validation methods to check their performance, we obtained around 94.99% accuracy, spending 0.06s on building the model.

Table 1: Random Forest Classifier Overall Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Split (70%)	43	2	95.55%	0.0363	0.1532	0.17
5 Fold CV	143	7	95.33%	0.037	0.1531	0.05
10Fold CV	142	8	94.66%	0.0408	0.1624	0.03
15Fold CV	142	8	94.66%	0.0385	0.1613	0.14
20Fold CV	143	7	95.33%	0.0379	0.1558	0.03

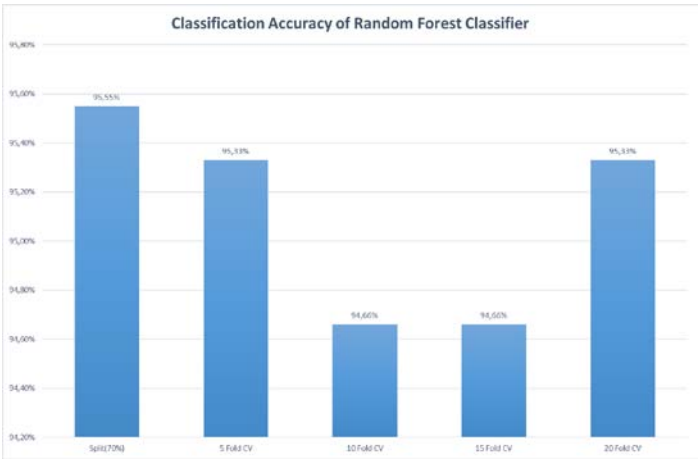

Figure 1: Classification Accuracy of Random Forest Classifier

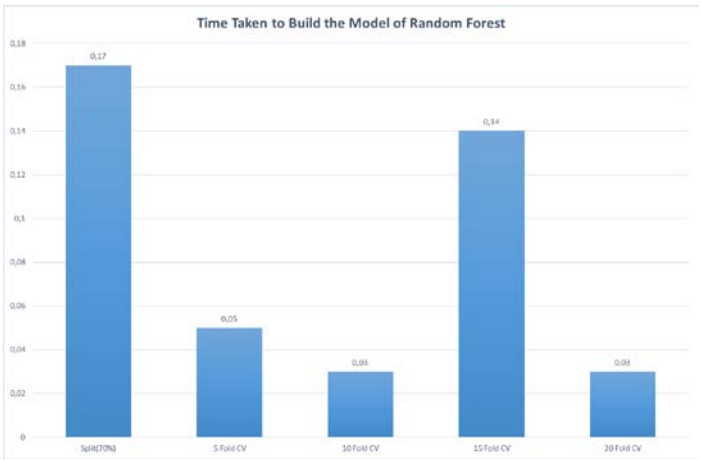

Figure 2: Time Taken to Build the Model of Random Forest Classifier

Table 2: Confusion Matrix – Random Forest Classifier (Split 70 %)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	14	0	0	14
Versicolor	0	16	0	16
Virginica	0	2	13	15
Predicted (Total)	14	18	13	45

Table 3: Confusion Matrix – Random Forest Classifier (5 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	50	0	0	50
Versicolor	0	47	3	50
Virginica	0	4	46	50
Predicted (Total)	50	51	49	150

Table 4: Confusion Matrix – Random Forest Classifier (10 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	50	0	0	50
Versicolor	0	47	3	50
Virginica	0	4	46	50
Predicted (Total)	50	51	49	150

Table 5: Confusion Matrix – Random Forest Classifier (15 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	50	0	0	50
Versicolor	0	47	3	50
Virginica	0	5	45	50
Predicted (Total)	50	52	48	150

Table 6: Confusion Matrix – Random Forest Classifier (20 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	50	0	0	50
Versicolor	0	47	3	50
Virginica	0	4	46	50
Predicted (Total)	50	51	49	150

b) Performance of J48Classifier

Table VII show the global evaluation summary of J48classifier using both of the test modes: splitting and different cross-validation methods. Fig.3 and Fig.4 display the performance of J48classifier in terms of classification accuracy and time taken on building the model. From Table VIII to Table XI we gave the confusion matrix for different test modes.

By applying these test modes using J48classifier we got 95.55% accuracy, spending 0.05s on building the model for the split mode. Using different cross-validation methods to check their performance, on average we obtained around 95.83% accuracy, spending 0.025s to build the model.

Table 7: J48 Classifier Overall Evaluation Summary

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Split (70%)	43	2	95.55%	0.0416	0.1682	0.05
5Fold CV	144	6	96%	0.035	0.1582	0.02
10Fold CV	144	6	96%	0.035	0.1586	0.02
15Fold CV	143	7	95.33%	0.0395	0.1758	0.03
20Fold CV	144	6	96%	0.0354	0.1586	0.03

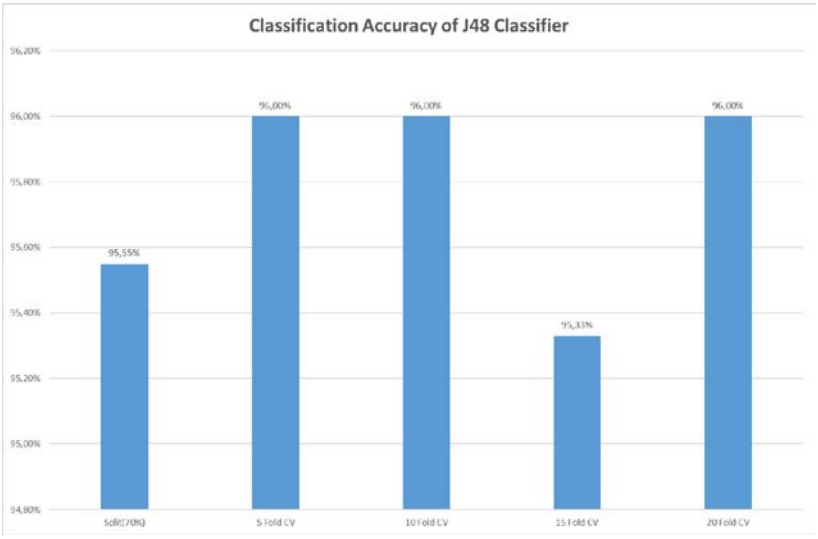

Figure 3: Classification Accuracy of J48 Classifier

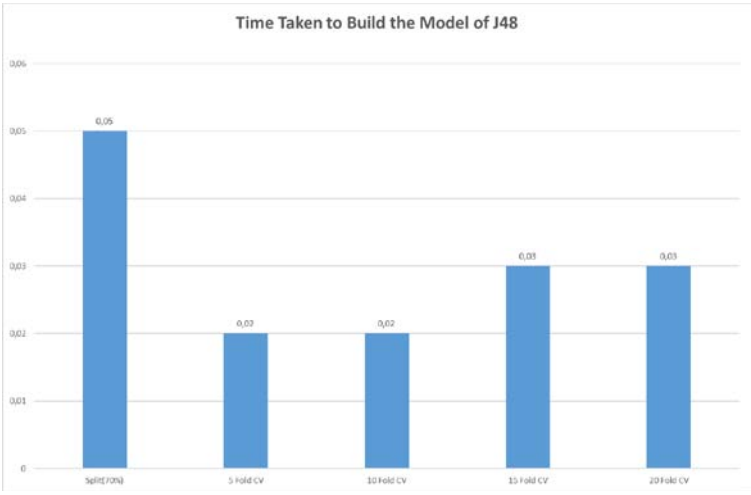

Figure 4: Time Taken to Build the Model of J48Classifier

Table 8: Confusion Matrix – J48 Classifier (Split 70 %)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	14	0	0	14
Versicolor	0	16	0	16
Virginica	0	2	13	15
Predicted (Total)	14	18	13	45

Table 9: Confusion Matrix – J48 Classifier (5 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	49	1	0	50
Versicolor	0	47	3	50
Virginica	0	2	48	50
Predicted (Total)	49	50	51	150

Table 10: Confusion Matrix – J48 Classifier (10 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	49	1	0	50
Versicolor	0	47	3	50
Virginica	0	2	48	50
Predicted (Total)	49	50	51	150

Table 11: Confusion Matrix – J48 Classifier (15 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	49	1	0	50
Versicolor	0	47	3	50
Virginica	0	3	47	50
Predicted (Total)	49	51	50	150

Table 12: Confusion Matrix – Random Forest Classifier (20 Fold CV)

	Setosa	Versicolor	Virginica	Actual (Total)
Setosa	49	1	0	50
Versicolor	0	47	3	50
Virginica	0	2	48	50
Predicted (Total)	49	50	51	150

VI. Comparison of Random Forest and J48 Classifiers

Fig. 5 and Fig. 6 illustrate a comparison between Random forest and J48 according to classification accuracy and time taken on building the model.

Through the comparison of the performance using training set (70%) process and various cross-validation methods between Random Forest and J48 classifiers depending on time taken on building the model, CA, MAE, and RMSE values, we reached thatJ48 classifier outperforms Random Forest.

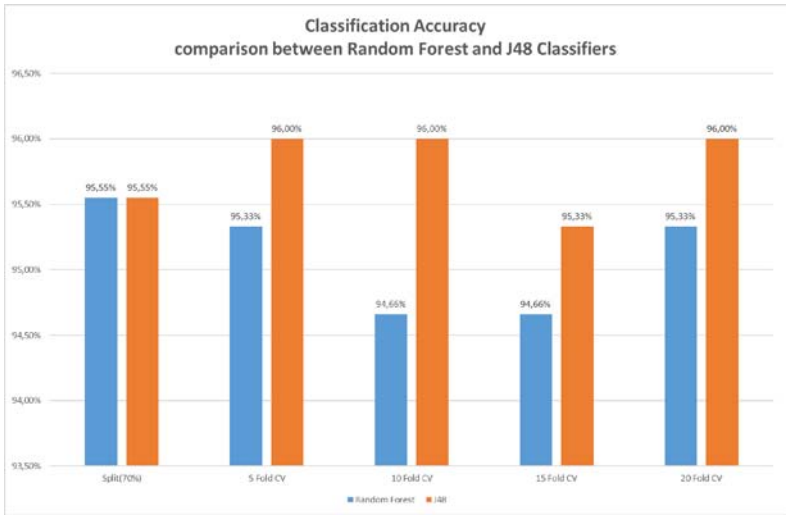

Figure 5: Classification Accuracy, Comparison between Random Forest and J48 Classifiers

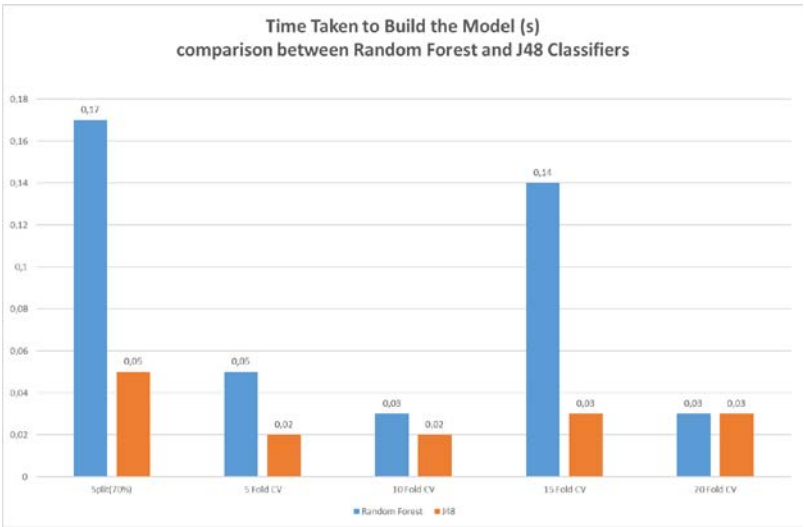

Figure 6: Time Taken to Build the Model, Comparison between Random Forest and J48 Classifiers

VII. Conclusion

This research work compares the efficiency of Random Forest and J48 Classifiers for IRIS variety prediction. The test is accomplished using WEKA 3.9in a machine with a processor i5-2430M 2.40 GHz and 4.00GB in RAM. Also, we compare the performance of both of the classifiers in terms of different scales of effectiveness evaluation. At last, we observed that J48classifier performs best than Random Forest classifier for IRIS variety prediction by taking different measures, including classification accuracy, Mean Absolute Error, and Time Taken to Build the Model.

References Références Referencias

1. Daniel T. Larose, Chantal D. Larose," An Introduction to Data Mining", Computer Science, 2014.
2. Mehmed M. Kantardzic," Data Mining: Concepts, Models, Methods, and Algorithms", Computer Science, 2002.
3. Margaret H. Danham, and S. Sridhar, "Data Mining, Introductory and Advanced Topics". Person Education, 1st Edition, 2006.

4. Uci Machine Learning Data Repository Http://Archive.Ics.Uci.Edu/Ml/Datasets/Iris.
5. Avci Mutlu, Tülay Yildirim "Micro Controller Based Neural Network Realization And Iris Plant Classifier Application", International Xii. Turkish Symposium on Artificial Intelligence and Neural Network,(2003).
6. Kavitha Kannan,. "Data Mining Reporton Iris And Australian Credit Card Dataset", School Of Computer Science And Information Technology, University Putra Malaysia, Serdang, Selangor, Malaysia, 2010.
7. Leo Breiman, "Random Forests". Machin Elearning. 45(1): 5-32, 2001.
8. P. Hamsagayathri Andp. Sampath," Decision Tree Classifiers for Classification of Breast Cancer". Int J Curr Pharm Res, Vol 9, Issue 2, 31-36, 2017.
9. Tina R. Patil, And S. S. Sherekar," 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.
10. Neeraj Et Al," Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June - 2013, Pp. 1114-1119.