IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. I (Jul - Aug 2018), PP 40-44
www.iosrjournals.org

Decision Support System Using Weighted Random Forest For Astronomical Data

Ms. Shilpa Gedam[1], Dr. Mrs. Ranjana Ingolikar[2]
[1] *(IICC, RTMNU Campus, Nagpur, Maharashtra, India)*
[2]*(Saint Francis De' Sales College, Seminary Hills, Nagpur, Maharashtra, India)*
Corresponding Author: Ms. Shilpa Gedam

Abstract: *Decision Support System plays an important role in making decisions. Decision support system may use data mining techniques for solving problem. Astronomy is an area where Data Mining has been playing a major role. As the astronomical data is very huge, the classification of celestial bodies is the main issue of concern. To improve the classification accuracy a new improved weighted random Forest algorithm is suggested. A decision support system is designed using Weighted Random forest algorithm. The algorithm is implemented in Java. It is observed that weighted random forest performs better than random forest and other tree based data mining classification techniques.*

Keywords: *Decision Support System, Ensemble learning, Random Forest, Weighted Random Forest*

Date of Submission: 02-07-2018 Date of acceptance: 21-07-2018

I. Introduction

Now a days a computer based programs can be used to assist the decision maker in taking right action. The computer aided program (software) that helps the decision maker is known as decision support system. The decision support system provides a scientific basis for each decision that is taken. Decision support system is developed keeping in view the requirements of the problem. In real life, decisions are normally taken with respect to the previous knowledge. So, the decision support system also uses the previous data that is available. In addition to this it also uses mathematical and scientific methods. The problem that uses previous data is called information analysis problem. And information analysis in other words is called Data mining. In other words, extracting knowledge from the available data in a scientific way is known as Data Mining. There are different ways (techniques) in which data can be extracted. Keeping in view different types of decision making problems, Thomas L. Saaty and Daji Ergu[1] discussed various decision making methods for handling conflicting and non conflicting criterion of 16 different categories of criterion. Jijun et al [2] presented a method of grey related analysis to handle multiple attribute decision making. They used interval fuzzy numbers for standardizing the input. Fiji Ran et al[3] proposed a novel multiclassifier ensemble method based on dynamic weights to increase decision accuracy and reliability. They defined an algorithm for decision credibility which describes the real time importance of the classifier by dynamically assigning fusion weight to the classifier.

In the present work deals with classification of astronomical data. Similar type of classification was performed on astronomical data using random forest by Franco-Arcega et al.[4], Honghai Wang [5] , Zhenping YI and Jingchang PAN [6]. The present work aims at checking the efficiency of ensemble based on decision tree (Random Forest [7]), generation of Weighted Random Forest Algorithm and comparative analysis of different tree based algorithms for classification. The paper is divided into five sections. Section data describes the data collection and generation for performing data mining. Section method discusses the weighted random forest algorithm. Section Experiment and results discuses the experimental results obtained using Weighted random forest and other tree based mining techniques and Section Conclusion discusses the conclusion of the work.

II. Data

The Slogan Digital Sky Survey(SDSS) is the largest optical survey of the astronomical bodies such as stars, galaxies, asteroids etc. SDSS contains data of ~10^9 objects(data release 9,10) covering 1/3 of the sky[8]. Slogan Digital Sky Survey (SDSS) is used as a data source for this astronomical data. The Slogan Digital Sky Survey maps one-quarter of the entire sky in detail, determining the position and absolute brightness of hundreds of millions of celestial objects. It also measures distances from the earth to more than a million galaxies and quasars [8]. The SDSS has obtained high resolution pictures of one quarter of the entire sky. The astronomical bodies are observed using an instrument called a spectrograph. The spectrograph generates spectra of each astronomical body which provides information about the astronomical body. SDSS uses different color filters (u, g, r, i, z) designed to let in light around a specific wavelength. Fig. 1 shows the image of star and Fig. 2 shows the spectra obtained using SDSS [9].

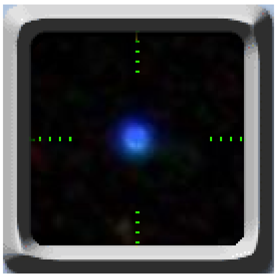


Figure 1. Image of star

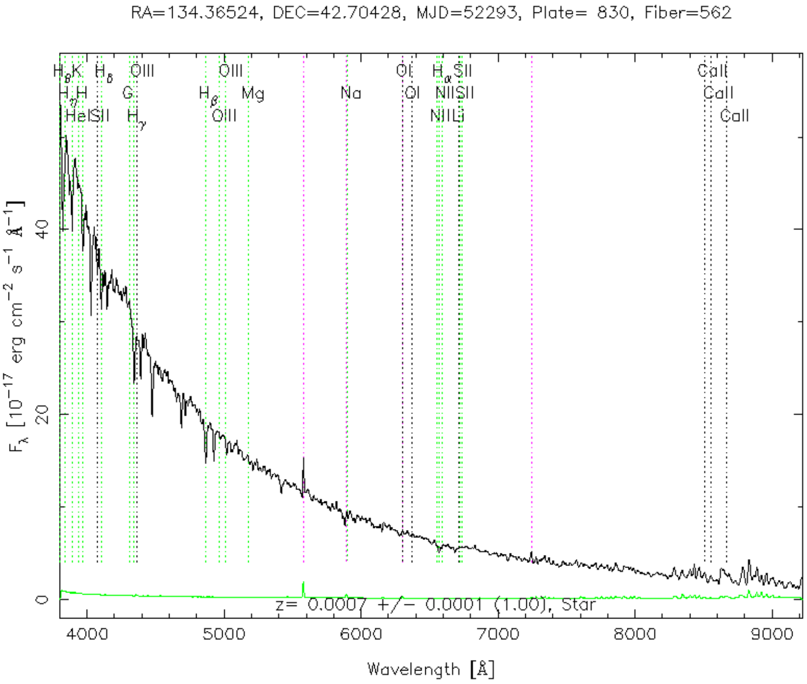


Figure 2. Spectra of a star

From the available spectra of the individual object right ascension value, declination value , colour , redshift value of the celestial object , u, g, r, i, z SDSS filter values, temperature, wavelength, intensity of light from the object , radial velocity[10] and class of star are recorded. From SDSS, spectra for 1500 such stars are obtained.

III. Method

The data is divided into training set (80 %) and testing set (20 %). The training set is divided into small data sets. Each small data set is again divided into indata and outdata. Tree is constructed using indata and tested using outdata and weight is generated. In the similar manner, trees are constructed from each data set, tested and weight label are generated. While building the model, weights are checked. If the weight label is Most accurate or Accurate then only that tree is taken for building the forest. So in this way less useful trees will be omitted and only useful trees will remain, which may help in giving better classification results and also number of trees used for building the random forest will be less. The model is built using training set and then testing of model is done using testing set. Weighted Random Forest algorithm is as follows

Weighted random Forest Algorithm: Part A gives the description of the algorithm and explanation of the terms used and Part B gives the Weighted Random Forest Algorithm.

- ***Part A:***

Binary Tree of size less or equal to 11 are generated and Weight for each tree is calculated.
Weight is generated using the formula
Weight= ((col1+col2+….+coln) /14) * 0.001…………………..(1)

Number 14 in the formula denotes the total number of attributes considered and 0.001 is used to normalized the weight value.
For Label generation using weights, sigmoidal membership function is used.
The weight value varies from 0 to 1. Table 1 shows the weights and corresponding labels given to trees. During model development, the threshold value for weight is taken as 0.75. The weight generated fits the sigmoidal membership function. Figure 3 shows the labels that are generated using sigmoidal function.

Table 1: Weights and Labels

Weights	**Labels**
1 - 0.90	Most accurate
0.89 - 0.75	Accurate
0.74 - 0.50	Less accurate
0.49 - 0	Least accurate

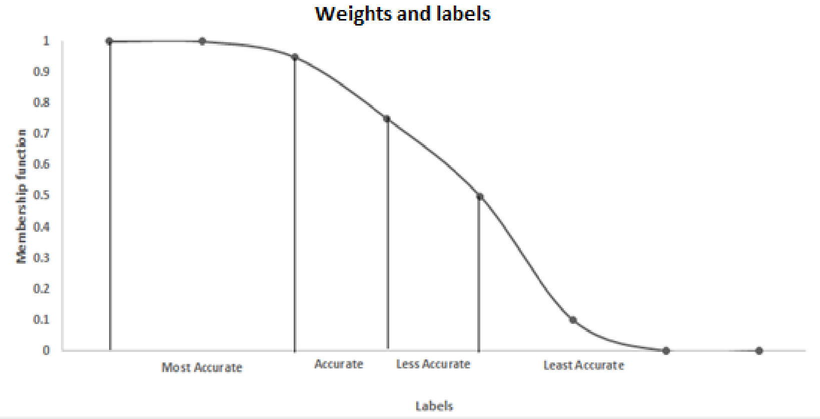


Figure 3: Labels using Sigmoidal function

- ***Part B:***

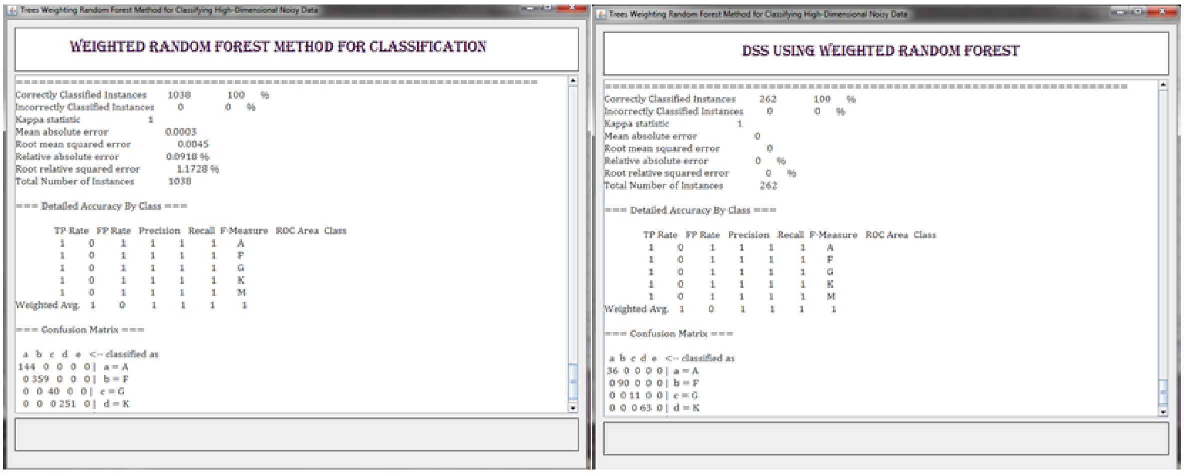
Let T1 denote be the training set,
S denotes the number of small data sets obtained from T1 and
T2 denotes the testing set.
Input:
T1: Training Set
Output:
W1: Weighted Random Forest
Algorithm : Weight_Random_Forest(T1)

```
1. For each small data set i = 1 to S
a)  Take 80% of records as indata and 20% of records as outdata
b)  Generate binary decision tree for indata for dataset with attributes A1, A2,…, An
c)  Generate Weight using column values col1, col2, …, coln and test the tree using outdata .

2. Take only trees whose weight label is most accurate or accurate.
3. Build Weighted random forest model , W1 using trees obtained in step 2 and Test the model using test set T2.
```

IV. Experiment And Results

Netbeans environment is chosen for implementation. Java language is used for coding. Using the Weighted Random Forest algorithm, Decision Support System is developed for classification of Astronomical data. According to Morgan-Keenan system stellar spectra are of 10 types. Each spectra obtained from SDSS shows the intensity of light for wavelengths in the region 3800 Å to 9200 Å. The stars are classified into the class types A, F, K, G, M. Some builtin classes of Weka [11] are used during implementation. The output screenshots of Decision Support System using Weighted Random forest for sample size 1300 is shown in Fig. 4. 80% data (1038 records) is taken as training Set and 20% data (262 records) data is taken as test data.


Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data
WEIGHTED RANDOM FOREST METHOD FOR CLASSIFICATION
Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data
DSS USING WEIGHTED RANDOM FOREST

```
=====================================================================
Correctly Classified Instances      1038          100    %
Incorrectly Classified Instances      0             0    %
Kappa statistic                       1
Mean absolute error                  0.0003
Root mean squared error                0.0045
Relative absolute error              0.0918 %
Root relative squared error            1.1728 %
Total Number of Instances            1038

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
            1         0          1          1        1          1        A
            1         0          1          1        1          1        F
            1         0          1          1        1          1        G
            1         0          1          1        1          1        K
            1         0          1          1        1          1        M
Weighted Avg.   1         0          1          1        1          1

=== Confusion Matrix ===

  a   b   c   d   e   <-- classified as
144   0   0   0   0 |  a = A
  0 359   0   0   0 |  b = F
  0   0  40   0   0 |  c = G
  0   0   0 251   0 |  d = K
```

```
=====================================================================
Correctly Classified Instances       262          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     %
Total Number of Instances            262

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
            1         0          1          1        1          1        A
            1         0          1          1        1          1        F
            1         0          1          1        1          1        G
            1         0          1          1        1          1        K
            1         0          1          1        1          1        M
Weighted Avg.   1         0          1          1        1          1

=== Confusion Matrix ===

 a  b  c  d  e   <-- classified as
36  0  0  0  0 |  a = A
 0 90  0  0  0 |  b = F
 0  0 11  0  0 |  c = G
 0  0  0 63  0 |  d = K
```

Figure 4. Output screenshots for training and test set (Sample size 1300)

For comparison different tree based mining methods are considered. They are Hoeffding Tree, J48, LMT, Random Tree, REPTree and Decision Stump. To study the behavior of each method, the data is taken in an incremental manner like 300 records, 500 records, 750 records, 1000 records, 1300 records and 1500 records. Weka [11] Tool is used for creating a model using different tree based data mining methods. The classification accuracy of random forest method is compared with other tree based mining method. For each sample 80% data is taken for building the model and 20% data is used for testing the model. Table 2 shows the performance of all tree based data mining methods (in percentage) for different sample sizes. E in the table 1 denotes the minor error.

Table 2: Classification performance of all tree based data mining methods for different sample sizes

Tree based Classification Method	Sample Size 300	 500	 750	 1000	 1300	 1500
Hoeffding Tree	96.61	17.00	44.00	49.26	47.71	49.51
J48	99.22	≅100.00 (E)	≅100.00 (E)	≅100.00 (E)	≅100.00 (E)	≅100.00 (E)
LMT	98.31	99.00	99.00	98.52	99.00	99.00
Random Tree	98.31	99.00	99.00	99.00	98.47	98.02
REPTree	99.17	99.00	99.00	96.06	99.00	99.67
Decision Stump	74.58	72.00	69.33	66.01	58.40	57.10
Random Forest	≅100 (E)	≅100(E)	≅100 (E)	≅100 (E)	≅100(E)	≅100 (E)
Weighted Random Forest (Proposed Method)	≅100 (E)	≅100 (E)	≅100 (E)	≅100 (E)	100.00	≅100 (E)

From table 2 , it can be seen that Random Forest and Weighted Random Forest gives ~100 % classification accuracy for all sample sizes but with error. The error that is generated is measured in the form of root mean square error. The comparison of Root mean square error generated using different tree based mining methods for sample sizes ranging from 300 to 1500 is shown in table 3.

Table 3: Comparison of Root Mean Square Error for sample sizes (300, 500, 750, 1000, 1300, 1500)

Tree based mining Method	Root Mean Square Error for Sample sizes					
	300	**500**	**750**	**1000**	**1300**	**1500**
Hoeffding tree	0.109	0.3778	0.3439	0.3184	0.3565	0.3538
J48	0.06	0.025	0.0452	0.1718	0.0353	0.0451
Decision Stump	0.2766	0.2785	0.832	0.2875	0.3139	0.3194
Random Tree	0.0823	0.0632	0.05132	0.1639	0.0781	0.091
Reptree	0.085	0.0214	0.0515	0.1251	0.0258	0.0227
LMT	0.081	0.0748	0.0821	0.0672	0.0912	0.0857
Random Forest	0.0197	0.0528	0.0403	0.0291	0.0205	0.0202
Weighted Random Forest (Proposed method)	0.028	0.0246	0.0076	0.003	0	0.0031

Fig. 4 shows the comparison of root mean square error for Random Forest and Weighted Random forest.

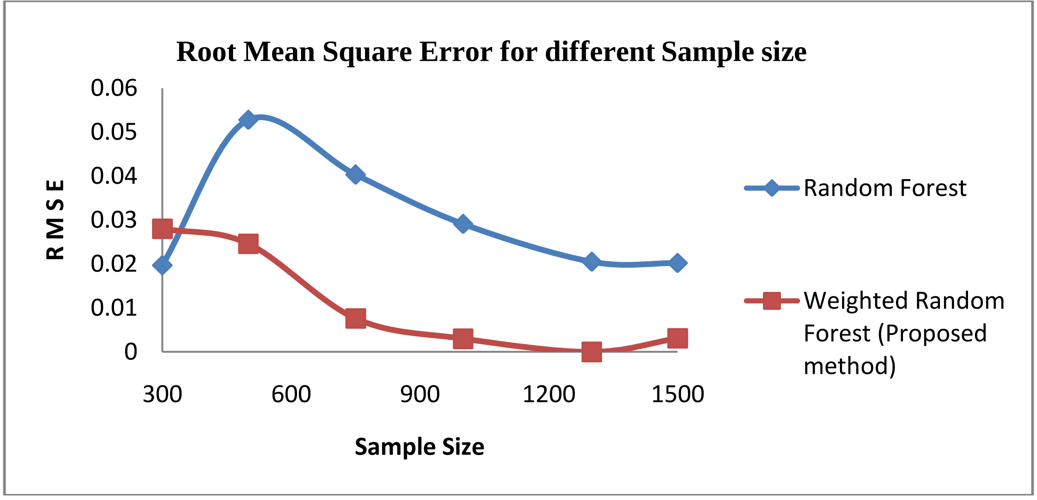


Figure 4: Comparison of Root Mean Square Error for different Sample Size using Random Forest and Weighted Random Forest

From Table 3 and Fig. 4 , it is observed that Weighted Random forest lowers the root mean square error for almost all sample sizes. Weighted random Forest shows its optimal performance with 0 root mean square error for sample size 1300.

V. Conclusion

Data mining techniques include ensemble learning which consists of many classifiers and helps in exact decision making process. A decision support system is developed using weighted random forest for astronomical data. Classification is performed on astronomical data to categorize star into different classes. Weighted Random forest showed best results for sample size 1300. Comparative analysis also gives best performance of Weighted random forest over other tree based data mining techniques.

References

[1]. Thomas L Satty and Daji Erdu, “When is a Decision-Making Method Trustworthy? Criteria for Evaluating Multi-Criteria Decision-Making Methods”, International Journal of Information Technology and Decision Making , Vol 14, Issue 6, Nov 2015.

[2]. Jijun Zhang, Desheng Wu and D. L. Olson, “ The method of grey related analysis to multiple attribute decision making problems with interval numbers”, Mathematical and Computer Modelling , Vol 42, Issue 9-10, 991-998, Nov 2005.

[3]. Fiji Ren, Yanqiu Li and Min Hu, “Multi-Classifier ensemble based on Dynamic Weights”, Multimed Tools Appl, Springer, 2017.

[4]. Franco-Arcega, L.G. Flores-Flores,Ruslan F. Gabbasov, ”Application of decision trees for classifying astronomical objects”,12th Mexican International Conference on Artificial intelligence,IEEE,181-186,2013.

[5]. Honghai Wang,”Pattern classification with random decision forest” International Conference on Industrial Control and Electronics Engineering,IEEE,128-130,2012.

[6]. Zhenping YI and Jingchang PAN,”Application of Random Forest to Stellar Spectral Classification”, 3rd International Congress on Image and Signal processing”, IEEE, 3129-3132, 2010.

[7]. Leo Beiman, random forest, machine Learning, 45, 5-32,2001.

[8]. D.G.York, et al., and SDSS Collaboration. The Sloan Digital Sky Survey : Technical Summary. AJ, 120:1579-1587, September 2000.

[9]. Astrophysical Research Consortium, Slogan Digital Sky Survey [online][Cited December 21, 2013.] http://skyserver.sdss.org/dr9.

[10]. Jayant Vishnu Narlikar, An Introduction to Cosmology, third Edition, Cambridge University Press, 2002.

[11]. M.Hall, E.Frank, G. Homes, B. Pfahringer, P.Reutemann and I.H. Witten. The weka data mining software: An update. SIGKDD Explorations, 11(1):10-18, 2009.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Ms. Shilpa Gedam. " Decision Support System Using Weighted Random Forest For Astronomical Data." IOSR Journal of Computer Engineering (IOSR-JCE) 20.4 (2018): 40-44.