RESEARCH REVIEW JOURNALS
NCIDCAS-2018
DEC- 2018
Special Issue

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Spectral Classification using DSS based on Weighted Random Forest and other mining techniques

[1]S.R. Gedam & [2]R. A. Ingolikar

[1]IICC, RTMNU, Nagpur, (India)
[2]Department of Computer Science, SFS College, RTMNU, Nagpur (India)

ARTICLE DETAILS

Article History
Published Online: 22 Dec 2018

Keywords
Ensemble learning; Weighted random forest; Decision Support System; Random forest.

***Corresponding Author**
Email: shilpamanke[at]gmail.com
r_ingolikar[at]rediffmail.com

ABSTRACT

Spectral classification is an area of great importance for most of the astrologers. Different classification methods have been used for this purpose. Ensemble learning is a classification method which uses different classifiers and result of classification is based on the vote obtained by individual classifier. A decision support system based on ensemble learning technique for spectral classification is devised. The tree based ensemble learning classifier is called weighted random forest classifiers. The classification results obtained using weighted random forest and different mining techniques are compared. It is observed that weighted random forest performs better than other mining techniques.

1. Introduction

We take decisions in day today life. Decisions are normally taken with the help of previous knowledge or by learned persons. Nowadays a computer aided can be used to assist the decision maker called Decision Support System. Every decision that is taken using decision support system has a scientific base to it. Decision support system can be designed to handle different categories of problems so D. Ergu and T. Saaty discussed various decision making methods for handling conflicting and non-conflicting criterion [1]. Jijun et al[2] presented a method for handling multiple attribute decision making. Fiji Ran et al [3] proposed an ensemble method based on dynamic weights to increase decision accuracy and reliability.

Classification problem is one of the type of problem which can be dealt with decision support system. Classification is information analysis (data mining) problem in which the objective is to predict a class name for the unknown data using the available data. Random forest uses an ensemble learning technique for classification. It is a tree based classifier which takes the help of different trees for predicting class name for the data. Present work deals with spectral classification. Similar astronomical classification was performed by Franco-Arcaga [4], Honghai Wang [5], Zhenping YI and Jingchang PAN [6]. The present work aims at making the comparative study of classification results of different data mining techniques. The paper is divided into five sections. The section Data describes the data used in the classification process. Section Method discusses the weighted random forest method and other mining techniques. Section Experiment and Results discusses the classification results. Section Conclusion discusses the conclusion of the present work.

2. Data

The Slogan Digital Sky Survey (SDSS) is the largest optical survey of the astronomical bodies (objects) including stars, galaxies, asteroids etc., and contains data of ~10^9 objects(data release 9) [7]. The data is available in the form of spectra or image. The images of the astronomical bodies are taken by SDSS in five photometric bands u, g, r, i and z in the optical wavelength range 0.3-1.0μm. These bands give enough information helpful to classify these objects as stars or other celestial body. The available spectra of the stars are observed and right ascension value, declination value, object's redshift, velocity, intensity of light, temperature is calculated. Figure 1 shows the image of the star obtained using SDSS Navigation tool and Figure 2 shows the sample spectra of a star.

According to Morgan-Kennan classification, stellar spectra were divided into totally 10 class types : O, B, A, F, G, K, M, R, S, N. We generated class type as the target output of the classification. The wavelength in the available spectra ranges from 3800 to 9200 Angstrom so our classifier classifies A, F, G, K, M class types. A sample size of 1500 is generated. Table I shows the sample data of 10 records.

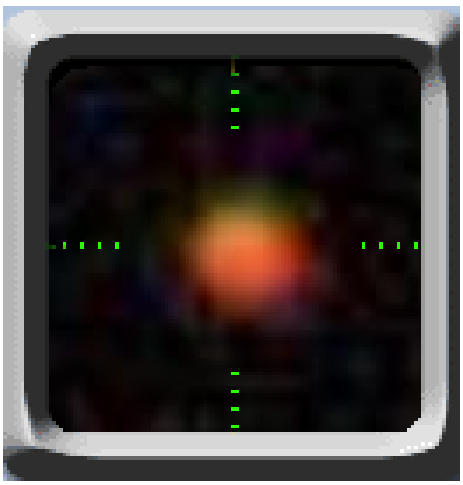

Figure 1. Image of the star

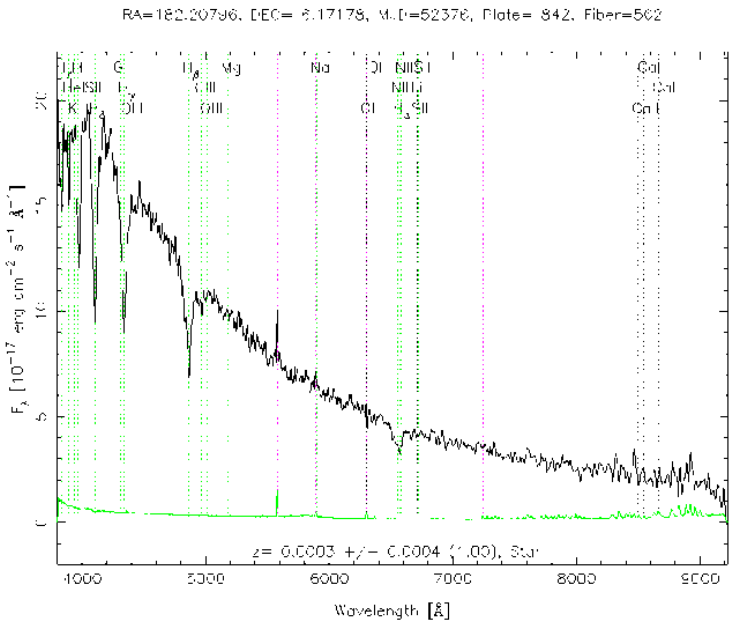

Figure 2. Sample Spectra of the star

TABLE I. Sample data of 10 records

S. no	RA	DEC	U	Intensity	Wave length	R. velo	Temp	Class
1	53.63515	-5.42961	24.35	7	7600	60	3812.86	K
2	42.69537	1.15301	15.19	170	4000	-90	7244.43	F
3	356.6797	16.0897	17.6	120	4000	-180	7244.43	F
4	134.3652	42.7043	17.91	50	3800	210	7625.72	A
5	182.208	6.17178	18.91	20	4100	90	7067.74	F
6	176.7317	1.15892	17.6	80	3800	60	7625.72	A
7	215.8062	0.42469	21.16	48	7700	150	3763.34	K
8	220.2851	1.17219	16.64	160	3800	30	7625.72	A
9	183.6272	1.08106	20.48	38	7600	-30	3812.86	K
10	195.0071	-1.17447	15.83	500	4500	-60	6439.49	F

3. Method

Classification is performed using Weighted Random Forest [8]. Weighted Random Forest uses decision trees for classification. During the training phase the weight value is generated for each tree. The weight value varies from 0 to 1. Weight value is generated using sigmoidal membership function. With the help of these weights the trees which are less useful are deleted and only useful trees are considered for model development. Figure 3 shows the working of the weighted random forest. The classification is performed with the help of Decision Support System. Java language is used for coding and Net beans environment is chosen for implementation of Decision Support System.

For comparative analysis different data mining techniques are used like BayesNet, NaiveBayes, Multilayer perceptron, KStar, Bagging, Multiclass Classifier, Decision Table and Random Forest.

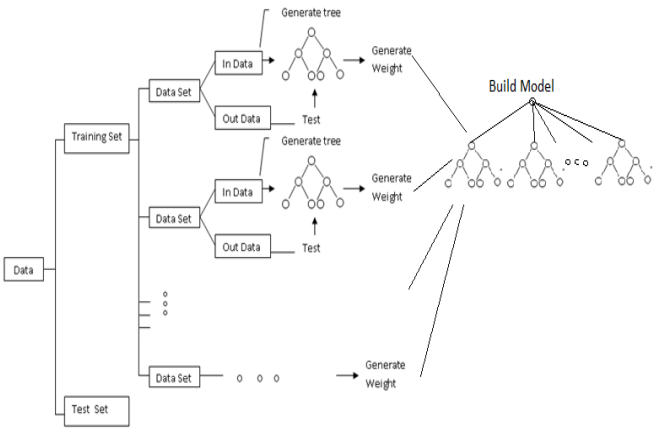

Figure 3. Working of the Weighted Random Forest

BayesNet is a classification method which assumes that all variables are discrete and finite. BayesNet treats the attributes and class as a random variable. The random variable is defined by a probability density function. The probability that

x object belongs to class C is calculated using probabitity density function P(C/x). This probability is determined using Bayes theorem [9].

Naïve Bayes algorithm is used for predictive modeling. It is collection of algorithms based on Bayes theorem. All the algorithms assume that every pair of feature being classified is independent of each other [10].

Perceptron is a pattern recognition machine. It has multiple inputs fully connected to an output layer. Multilayer Perceptron extend the perceptron with hidden layers of processing elements that are not connected directly to the external world. Multilayer Perceptron is a type of deep learning network [11].

KStar is an instance based classifier that is the class of a test instance is based upon the class of those training instances similar to it. This similarity is determined by some similarity function. Sometimes it may differ from other instance based learner by selecting different function such as entropy based distance function [12].

Bagging is also called Bootstrap Aggregation. Bagging is an ensemble method that combines the predictions from multiple machine learning algorithms to make more accurate prediction than an individual model. This procedure is used to reduce the variance for those algorithm that have high variance [13].

Multiclass Classifier is type of supervised machine learning. It uses Decision tree (data is visualized in the form of tree), Support vector machine (feature vector is high dimensional) and K nearest Neighbour (does not depend on structure of data) classifiers on the training data to predict the label for the test data [14].

Decision tables are classification models used for prediction. They are induced by machine learning algorithms. A decision table consists of a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attributes to form another table [15].

Random forest is a recently proposed ensemble method [16] which uses many tree classifiers and aggregates their results. Random forest uses different bootstrap sample of data to construct each tree. Then a subset of predictors is chosen randomly, each node of the trees is split using the best among the subset instead of all predictors [17]. There are several ways to calculate output of random forest. The simplest is simple majority voting method for classification, while average output of trees is considered.

4. Experiment and Results

For sample size 1500 experiment is performed. 80% of sample size is taken as training data and 20% as test data. Weka Software[18] is used for generating the model for classifiers BayesNet, NaiveBayes, Multilayer perceptron, KStar, Bagging, Multiclass Classifier, Decision Table and Random Forest. The classification result and root mean square error (RMSE) generated during the classification are used for measuring the efficiency of the model.

Table II shows the classification performed for the training data and Table III shows the classification results for the test data.

Table II . Classification result for the training data

Mining Technique	RMSE	Classification Result (%)
Bayes Net	0.0177	99.9165
Naïve Bayes	0.0899	98.4962
Multilayer Perceptron	0.0561	99.3317
K Star	0.0231	99.6658
Bagging	0.0332	99.6658
Multi Class Classifier	0.353	99.9967
Decision Table	0.0107	99.9025
Random Forest	0.0210	≅ 100
Weighted Random Forest (Proposed Method)	0.0058	≅ 100

Table III. Classification result for the test data

Mining Technique	RMSE	Classification Result (%)
Bayes Net	0.0330	99.9261
Naïve Bayes	0.0647	98.6799
Multilayer Perceptron	0.0484	99.3398
K Star	0.1214	95.7096
Bagging	0.0202	99.988
Multi Class Classifier	0.3532	99.3399
Decision Table	0.0108	99.9817
Random Forest	0.0202	≅ 100
Weighted Random Forest (Proposed Method)	0.0031	≅ 100

As the classification of data is ≅ 100 % for Random Forest and Weighted Random Forest. The efficiency of both Random forest and Weighted Random forest is checked by varying the sample sizes. For different sample sizes both random forest and weighted random forest has shown ≅ 100 % accuracy. Table IV shows the detailed results of classification obtained by varying the sample size. Figure 4 shows the comparison of RMSE for Random Forest and Weighted Random Forest.

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

Sample Size	Random Forest	Weighted Random Forest
300	0.0197	0.028
500	0.0528	0.0246
750	0.0403	0.0076
1000	0.0291	0.003
1300	0.0205	0.000
1500	0.0202	0.0031

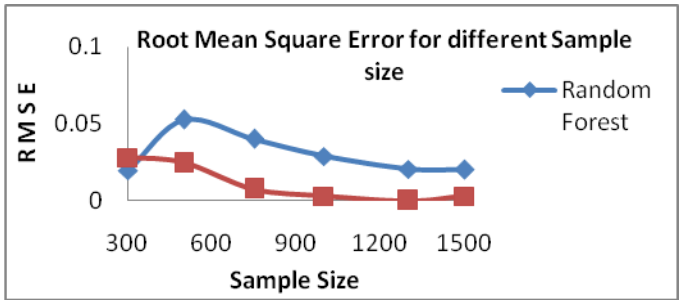

Figure 4. Comparison of Root Mean Square Error for different Sample Size using Random Forest and Weighted Random Forest

From Figure 4, it is clearly seen that the Weighted Random forest shows the best performance when sample size is 1300 with 100% classification accuracy and zero root mean square error.

5. Conclusion

In this study, we performed spectral classification using different mining techniques. The classification results obtained are compared. The results show that ensemble learning is a better method than individual classifier. Random Forest Method is an effective method in addition to it, Weighted Random Forest (Proposed Method) also performed better. Weighted Random Forest has shown its best performance for sample size 1300. As future work, more attributes that describe the astronomical objects can be considered and other celestial bodies can also be classified for large volumes of data.

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 Electron ics 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. D.G. York, et al., and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. AJ,120:1579-1587, September 2000.
8. S. Gedam, R. Ingolikar," Decision Support System Using Weighted Random Forest For Astronomical Data", IOSR Journal of Computer Engineering ,Volume 20, Issue 4, Ver. I (Jul - Aug 2018), PP 40-44.
9. Introduction to Bayes Net? (1995). Retrieved 10 20, 2018, from Tutorial on Bayes Network with Netica: http://www.norsys.com/Sec_A/tutA1.htm
10. Naive Bayes. (2007). Retrieved 9 21, 2018, from Scikit learn developers: http://scikit-learn.org/stable/module/naive_bayes.html
11. Edu, C. (2008). Chapter3 Multilayer perceptron. Retrieved 12 4, 2017, from CNEL: http://www.cnel.ufl.edu/courses/EEL6814
12. K-Star. (2008, 12 5). Retrieved 11 10, 2017, from KStar pentahoData Mining-pentaho Wiki: http://www.wiki.pentaho.com/display/DATAMINING/KStar
13. Bagging. (2016, 4 22). Retrieved 10 23, 2018, from Bagging and random Forest Ensemblle Algorithms for Machine Learning: https://machinelearningmastery.com/bagging-and-randon-forest-ensemble-algoritms
14. Multiclass classifier. (2010). Retrieved 11 6, 2017, from Multiclass classification using scikit-learn-GeeksforGeeks: https:www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/
15. Becker, B. (1998). Decision Table Classifier. Retrieved 5 30, 2018, from Visualizing Decision Table Classifier- ACM Digital Library: https://dl.acm.org/citation.ctm?id=721218
16. Leo Breiman, Random Forests, Machine Learning,45,5-32,2001.
17. Liaw, A.; Wiener, M. Classification and regression by randomForest, R News, 2:18-22,2002.
18. 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.