Study of Classification Models using Ensemble based and Non Ensemble based Mining techniques using Astronomical Data

S. R. Gedam*[1], R. S. Gedam[2], R. A. Ingolikar[3]
[1]*Department of Computer Science, Shivaji Science College, Congress Nagar, Nagpur*
[2]*Department of Physics, Visvesvaraya National Institute of Technology, Nagpur*
[3]*Department of Computer Science, Saint Francis De' Sales College, Seminary Hills, Nagpur*
[1]*shilpagedam2020@gmail.com*, [2]*rupeshgedam411@gmail.com*,
[3]*ranjana.ingolikar@gmail.com*

Abstract: *Accurate classification of huge data is a matter of concern for Data miners. In this paper, study of various data mining models for classification of astronomical data is done. Ensemble based and non-ensemble based methods are used for classification. Summary of all classification results is presented. Comparative analysis of classification results of Ensemble based and non-ensemble based classifier is done. The result shows that classification accuracy of Ensemble based classifier is better than non-ensemble based classifier .*

Keywords: **Ensemble based Mining, Non-Ensemble based Mining, Random forest, Weighted Random forest, Filtered classifier**

1. INTRODUCTION

Classification is troublesome if the data is huge. Classification model can be prepared using ensemble based and non-ensemble based methods. Ensemble based method considers the results of various individual classifiers and then aggregates the outcome for better classification results. Non Ensemble based methods are based on single classifier. Astronomical data (related to celestial bodies) being massive in number is considered here for classification.

Similar work related to astronomical data is done by SergiiKhlamov et al. used collection of light technology software for processing astronomical information. They also described the benefits of Online Data Analysis System for solving data mining problem[1] .TheeranaiSangjan et al presented a data level approach to solve imbalance data problem. They used classifiers like K-Nearest Neighbor, decision tree and Support Vector Machine for Investigation on a data set of Light curve profile[2]. M. Klush proposed a novel hybrid neural network approach for fully automated spectral and luminosity classification of stars. Hybrid neural system used neural classifiers and a semantic networks for similarity based reasoning [3]. Shiyu Deny et al applied Manhatten Distance density algorithm to variety of spectral data and concluded that the average classification stable number of the Manhatten Distance Density algorithm is smaller and the computing time is shorter [4]. Liangping Tu et al used local mean based K- Nearest Neighbor method for automated classification of Galaxies and Quasars. Their experimental results showed that local mean based K- Nearest Neighbor performs best and better than KNN [5]. Zhenping YI et al tried to evaluate the effectiveness of random Forest on stellar spectra. Their results also showed that random forest gave better efficiency and less root mean square error as compared to Multilayer perceptron network [6]. Jiang Bin et al presented a novel technique for automatically classifying massive stellar spectra selected from SDSS. Their results indicated the classification accuracy upto 90% [7]. In an attempt to classify

*S. R. Gedam

Celestial body we are dealing with Astronomical data (classification of celestial object-Star).

2. DATA

The astronomical data is generated using the spectra available on Slogan Digital Sky Server (SDSS)-10 [8]. SDSS provides information about various celestial bodies in the sky. SDSS provides data of about ~10^9 objects in the sky. Parameters that are considered for classification of Star are right ascesion of star, declination of star, intensity of light from star, wavelength of light, radial velocity of star, redshift of an object, temperature and colour of star. Classification model identifies the class of the star (using training and test data). The star is classified as of class A, F, K, G and M. Table 1 shows the sample data of size 20 records.

Table 1. Sample size of 20 records

Sr no	RA	DEC	u	g	r	i	z	Redshift	Intensity	Wavelength	Colour	Radial velocity	Temperature	Class
1	53.63515	-5.42961	24.35	22.44	20.34	19.14	18.49	0.0002	7	7600	RED	60	3812.86	K
2	42.69537	1.15301	15.19	14.01	13.47	13.81	13.42	-0.0003	170	4000	VIOLET	-90	7244.43	F
3	356.6797	16.0897	17.6	16.7	16.42	16.32	16.26	-0.0006	120	4000	VIOLET	-180	7244.43	F
4	134.3652	42.7043	17.91	18.25	18.74	19.08	19.37	0.0007	50	3800	UV	210	7625.72	A
5	182.208	6.17178	18.91	18.93	19.3	19.52	19.62	0.0003	20	4100	VIOLET	90	7067.74	F
6	176.7317	1.15892	17.6	16.7	16.42	16.32	16.26	0.0002	80	3800	UV	60	7625.72	A
7	215.8062	0.42469	21.16	18.46	17.1	16.48	16.19	0.0005	48	7700	RED	150	3763.34	K
8	220.2851	1.17219	16.64	16.84	17.29	17.6	17.88	0.0001	160	3800	UV	30	7625.72	A
9	183.6272	1.08106	20.48	18.14	16.98	16.51	16.25	-0.0001	38	7600	RED	-30	3812.86	K
10	195.0071	-1.17447	15.83	14.69	14.3	14.16	14.12	-0.0002	500	4500	VIOLET	-60	6439.49	F
11	239.6784	1.19479	17.6	16.7	16.42	16.32	16.26	0.0002	40	5600	GREEN	60	5174.59	K
12	241.4534	1.10398	18.85	17.58	17.08	16.86	16.77	0.0005	32	3800	UV	150	7625.72	A
13	255.6053	64.7947	17.8	16.64	16.13	15.9	15.78	-0.0008	110	4700	BLUE	-240	6165.47	F
14	247.2433	1.13857	22.21	20.29	19.44	19.11	18.87	-0.0001	28	5600	GREEN	-30	5174.59	K
15	24.3607	1.24467	20.06	19.15	18.73	18.56	18.51	0.0007	17	3800	UV	210	7625.72	A
16	30.3438	1.15482	17.97	16.90	16.54	16.43	16.41	0	96	4500	VIOLET	0	6439.49	F
17	48.25589	1.09948	19.15	18.73	18.67	18.74	18.84	0	22	3900	UV	0	7430.18	F
18	220.2851	1.17219	16.64	16.84	17.29	17.6	17.88	0.0001	160	3800	UV	30	7625.72	A
19	114.4405	38.8352	21.53	20.29	20.27	20.29	20.21	0.0005	6	3800	UV	150	7625.72	A
20	46.63369	-6.64844	18.34	17.89	17.81	17.81	17.87	0.0001	48	3800	UV	30	7625.72	A

3. METHODS

Classification Models are prepared using Weka [9] using ensemble and non-ensemble based methods. Non-ensemble based methods considered are BayesNet, Naïve Bayes, Logistic, SMO, KStar, LWL, MultiClass Classifier, Filtered Classifier, InputMapped Classifier, Jrip and ZeroR.

BayesNet is a classification method which assumes all variables to be 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 [10].

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 [11].

Logistic Regression classifier only supports binary classification problem. It has been adapted to support multiclass classification problem. It predicts a coefficient for each input value which is combined into a regression function and is converted using logistic function[12].

SMO stands for Sequential Minimal Optimization. It uses a specific algorithm used by Support Vector Machine. SMO works on numerical input values. It works by finding a line that separates the data into groups. SMO uses the instances in the training dataset that are closest to the line and separates the dataset into classes [13].

KStar is an instance based classifier. 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 [14].

Locally Weighted Learning uses an instance based algorithm to assign instance weight which are then used for classification. A classification is obtained from Naïve Bayes model by taking the attribute value of the test data as input. It can be used for classification or regression [15]. The subset of data used to train each locally weighted naïve Bayes model are determined by a nearest neighbor algorithm [16].

Filtered Classifier filters the dataset or alter it in someway like deleting a particular attribute, removing misclassified instances etc. For classification it selects any arbitrary classifier but initially the data is passed through a filter [19].

InputMapped Classifier addresses the incompatibility between training data and test data. It does this by building a mapping between the model built using training data and incoming test instances. If some important attributes are not found in the incoming instance then this classifier puts some nominal attribute value which the classifier has not seen before and then the model developed is trained accordingly for proper classification [20].

Jrip is propositional rule learner which does repeated incremental purning for error reduction. This classifier is proposed by Willaim W. Jrip divides the data into classes and generates rules by including all attributes of the class and for each instance until all the classes have been covered [21].

ZeroR is the classification method which relies on the target and ignores all predictors. ZeroR classifier just predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [22].

Ensemble based classification methods considered are Bagging, Multiclass classifier, Random Forest and Weighted Random Forest.

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 [17].

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 [18].

Random forest is a recently proposed ensemble method [23] 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 [24]. 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. **Weighted Random forest** is an updation of Random forest. Here forest is generated based on the weights assigned to trees. Weights are assigned in such a way that only useful trees are selected for model development [25].

4. Experimental Results

Classification Models are prepared for different sample sizes using ensemble and non-ensemble based methods. Sample Sizes are taken in an incremental manner. Initially all non-ensemble based methods are considered.

Table 2 and Table 3 shows the classification accuracy obtained using different non-ensemble based methods for Training data and Test data respectively.

Table 2. Classification Accuracy using non-ensemble based methods for training data

Sample Size / Methods	300	500	750	1000	1300	1500
BayesNet	100	100	100	100	99.904	99.917
Naïve Bayes	96.68	98.75	95.83	96.76	95.568	98.496
Logistic	100	100	100	100	100	100
SMO	89.21	94.25	96.17	96.63	95.279	95.489
KStar	100	100	100	100	99.904	99.666
LWL	95.07	88.25	87.17	88.15	82.081	94.152
Filtered Classifier	100	100	100	100	100	100
Input mapped Classifier	50.85	47.25	47.25	43.14	42.857	31.913
Jrip	100	99.75	99.75	99.88	99.904	99.917
ZeroR	50.21	47.25	47.25	43.14	34.584	31.913

Table 3. Classification Accuracy using non-ensemble based methods for test data

Sample Size / Methods	300	500	750	1000	1300	1500
BayesNet	98.3051	100	100	100	100	100
Naïve Bayes	88.1356	99	93.33	94.581	96.947	98.68
Logistic	100	96	97.33	95.567	98.473	98.68
SMO	86.4407	95	93.33	92.611	95.802	96.04
KStar	79.661	90	91.33	95.074	95.038	96.04
LWL	81.3559	88	87.33	66.997	82.06	96.37
Filtered Classifier	98.3051	100	100	100	100	100
Input mapped Classifier	50.8475	47	46	42.857	34.351	32.013
Jrip	100	99	99.33	100	100	100
ZeroR	50.8475	47	46	42.857	34.351	32.013

Table 4 shows the root mean square error generated by different Non ensemble based classifiers for different sample sizes.

Table 4.Root mean square error generated by different Non ensemble based classifiers

Non-Ensemble based Classifier	Sample Size					
	300	500	750	1000	1300	1500
Bayes Net	0.0925	0.0077	0.0016	0.015	0.0037	0.0015
Naïve Bayes	0.1802	0.0717	0.1421	0.1299	0.0972	0.0647
Logistic	0.0003	0.1216	0.0916	0.1332	0.0781	0.0705
SMO	0.3247	0.3183	0.3197	0.3197	0.3164	0.316
K Star	0.2768	0.1861	0.1643	0.1399	0.1239	0.126
LWL	0.2527	0.2183	0.2192	0.3033	0.2167	0.1756
Filtered Classifier	0.0823	0.005	0.002	0.008	0.002	0.002
Input mapped Classifier	0.3036	0.3696	0.3724	0.3758	0.3867	0.3887
Jrip	0.001	0.0633	0.0517	0.0012	0.001	0.0009
ZeroR	0.3633	0.3696	0.3724	0.3758	0.3867	0.3887

Figure 1 shows the Average Root Mean Square Error generated by different Non ensemble based classifiers during classification.

Now all ensemble based methods are considered. Table 5 and 6 the classification accuracy obtained using Random Forest and Weighted Random Forest for different

sample sizes using training data and test data respectively. Table 7 gives the root mean square error generated by each ensemble based method for different sample sizes during model development.

Table 5. Classification Accuracy using Ensemble based methods for training data

Ensemble based Classification Method	Sample Size					
	300	500	750	1000	1300	1500
Bagging	100	99.25	99.25	99.38	99.711	99.666
Multiclass Classifier	100	100	100	100	100	100
Random Forest	100	100	100	100	100	100
Weighted Random Forest	100	100	100	100	100	100

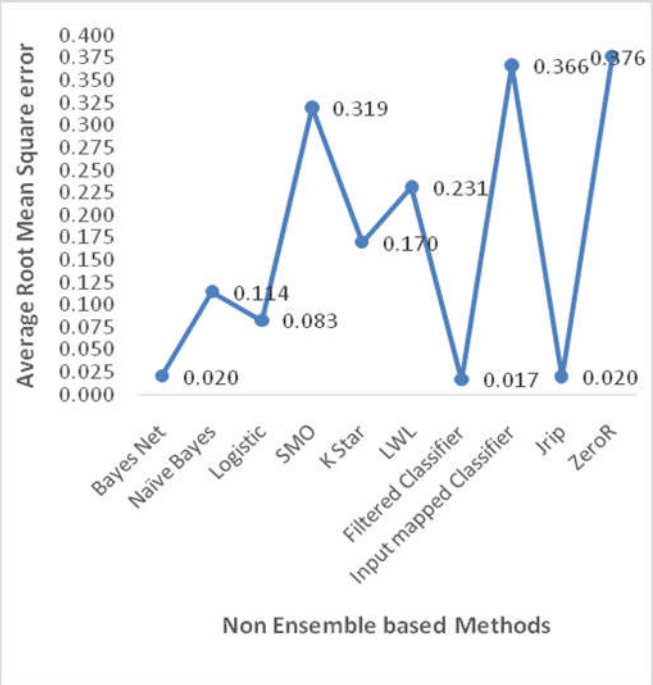

Figure 1.Average Root Mean Square Error generated by different Non ensemble based classifier

Table 6. Classification Accuracy using Ensemble based methods for test data

Ensemble based Classification Method	Sample Size					
	300	500	750	1000	1300	1500
Bagging	100	100	100	96.059	100	100
Multiclass Classifier	96.6102	99	97.33	98.03	99.618	99.34

Random Forest	100	100	100	100	100	100
Weighted Random Forest	100	100	100	100	100	100

Figure 2 given below shows the Average Root Mean Square Error generated by different Ensemble based classifiers during classification.

Table 7: Root mean square error generated during classification by ensemble methods

Ensemble based Classification Method	Sample Size					
	300	500	750	1000	1300	1500
Bagging	0.0326	0.02	0.509	0.1125	0.0034	0.0202
Multi Class Classifier	0.3533	0.3541	0.3543	0.3548	0.3531	0.3532
Random Forest	0.0197	0.0528	0.0403	0.0291	0.0205	0.0202
Weighted Random Forest	0.028	0.0246	0.0076	0.003	0	0.0031

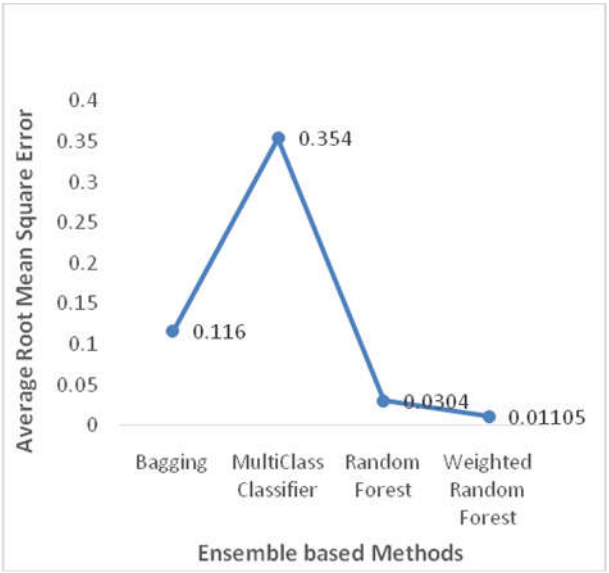

Figure 2. Average root mean Square error generated for different Ensemble based classifiers

From Figure 1, it is seen that Filtered classifier (Non Ensemble based Classifier) gives lowest average root mean square error (0.016883). From Figure 2, it can be concluded that Weighted Random Forest gives lower average root mean square error (0.0115) .

If both Filtered Classifier and Weighted Random forest are compared based on average root mean square error then it can be concluded that the performance of Weighted Random forest is better . So the performance of Ensemble based classifier is better as compared to non ensemble based classifiers.

5. Conclusion

In this paper study of different ensemble based and non ensemble based classifier is done. Classification Models are prepared using ensemble based and non ensemble based classifier . For comparative analysis, Root mean square error and average root mean square error generated during model development are considered. After comparative analysis, it can be concluded that the performance of Filtered classifier is best, if Non Ensemble based classification methods are considered and Weighted Random forst is best in case of ensemble based classification methods. Overall if Ensemble based and Non – Ensemble based methods are considered altogether than Ensemble based methods outshines in making better classification.

REFRENCES

[1] *SeriiKhlamovet al., "Colitec Software for Astronomical Data Sets Processing", Proceeding of IEEE second International Conference on Data Stream mining and processing(2018) Aug 21-25, 227-230.*

[2] *TheeranaiSangjan et al.."Classification of Astronomical objects Using Light Curve Profile."Proceeding of IEEE Eurasia Conference on IOT, Communication and Engineering(2019).*

[3] *M. Klush, "HNS- a hybrid neural system and its use for classification of stars", Proceeding of International Conference on Neural Network (1993).*

[4] *Shiyu Deny, Liangping Tu. "Classification of Celestial spectral based on improved desnsity clustering", Proceeding of 10th International Conference on Image and Signal Processing Bio Medical Engineering and Informatics. (2017).*

[5] *LiangpingTu, Huiming Wei and Liya Ai. "Galaxy and Quasar classification based on local mean- based K-Nearest Neighbor method", Proceeding of IEEE 5th International Conference on Electronics Information and Emerging Communication. (2015).*

[6] *Zhenping YI, Jingchang PAN. " Application of Random Forest to stellar spectral classification", Proceeding of 3rd International Congress on Image and Signal Processing , Volume 7. (2010).*

[7] *Jiang Bin, Wang Wenyu, Ma Chunyu, Wang Wei, Qu Meixia. "The Application of automative classification of massive SDSS spectra " , Proceeding of 2nd IEEE International Conference on Computer and Communication (2016) (1376-1380).*

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

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

[10] *Introduction to Bayes Net? Retrieved 10 20, 2018,from Tutorial on Bayes Network with Netica: http://www.norsys.com/Sec_A/tutA1.htm(1995).*

[11] *Naive Bayes. Retrieved 9 21, 2018, from Scikit learn developers: http://scikit-learn.org/stable/module/naive_bayes.html (2007).*

[12] *https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/*

[13] *https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/*

[14] *K-Star. Retrieved 11 10, 2017, from KStarpentahoData Mining-pentaho Wiki: http://www.wiki.pentaho.com/display/DATAMINING/KStar (2008, 12 5).*

[15] *https://wiki.pentaho.com/display/DATAMINING/LWL.*

[16] *Eibe Frank, Mark Hall, Bernhard Pfahringer. 2003. Locally Weighted Naive Bayes. In Proceeding of 19th Conference in Uncertainty in Artificial Intelligence(249-256), 2003.*

[17] *Bagging. Retrieved 10 23, 2018, from Bagging and random Forest Ensemblle Algorithms for Machine Learning: https://machinelearningmastery.com/bagging-and-randon-forest-ensemble-algoritms (2016, 4 22).*

[18] *Multiclass classifier. Retrieved 11 6, 2017, from Multiclass classification using scikit-learn-GeeksforGeeks: https:www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/ , (2010).*

[19] *https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781787122536/4/ch04lvl1sec46/classifying-unseen-test-data-with-a-filtered-classifier*

[20] *https://nodepit.com/node/org.knime.ext.weka37.classifier.WekaClassifierNodeFactory%23InputMappedClassifier%20(3.7)*

[21] *https://nodepit.com/node/org.knime.ext.weka37.classifier.WekaClassfierNodeFactory%23JRip%20(3.7)*

[22] *https://www.saedsayad.com/zeror.htm*

[23] *Leo Breiman. "Random Forests, Machine Learning", (2001), 45,(5-32).*

[24] *Liaw, A, Wiener, M. "Classification and regression by randomForest, R News", 2:18-22(2002).*

[25] *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 (2018). Pp 40-44.*