ISSN: 2321-7782 (Online)
Volume 3, Issue 9, September 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Application of Mining Techniques to Classification of Star

S. R. Gedam[1]
Inter Institutional Computer Centre
RTM Nagpur University
Nagpur(MS), India

Dr. R. A. Ingolikar[2]
Associate Professor, Department of Computer Science
S.F.S College, Seminary Hills.
Nagpur(MS), India

Abstract: ***Data mining techniques are used to analyze and discover correlations already present in databases. These techniques are very reliable and useful especially when large volumes of data are processed. These techniques have been applied to many areas such as marketing, medicine, diagnosis, business, biology, astronomy and others. In particular, astronomy requires techniques that allow the recognition or classification of astronomical objects from database that contain million of objects. Due to this, astronomers often deal with the analysis of large amounts of data from telescopes, seeking for several characteristics for their interpretation. Random forest is one of the useful techniques in data mining. From the results, it shows that ensemble learning is an effective classification method.***

Keywords: Random forest, Decision trees, Classification, Ensembles

I. INTRODUCTION

Data mining is an important area in the information analysis. It is defined as an information extraction activity whose goal is to discover hidden facts contained in databases [1]. Tasks of Data Mining can be classified into two categories :descriptive and predictive . Descriptive mining tasks characterize the general properties of the data using clustering, summarization, association rules or sequence discovery techniques. On the other hand, predictive mining tasks perform inference on the current data in order to make predictions [2] applying classification, regression or time series analysis techniques. These tasks help to solve several problems in different areas, such as medicine, industry, education, security, astronomy and many more[3].

Astronomy is an area where Data Mining has been playing a big role. Several techniques of Data Mining have been used to solve tasks in Astrology. Some of them are : an application of Bayesian analysis to the problem of star formation in young galaxies[4]; a Bayesian Markov Chain Monte Carlo method to determine whether the stars in the galaxies form in one monolithic collapse of a giant gas cloud, or if they form in a hierarchical fashion ; the use of computer vision and artificial neural network [5] in an application that classifies large number of galaxies which show up in the thousands of digitized images from sky surveys. Other works are the use of support vector machines [6] to explain the determination of the photometric redshift estimate for distant galaxies, and the use of a decision tree for classifying spatial data streams using a data structure called Peano Count Tree[5].

In this paper we focus on classification of stars. Classification is the process of finding a set of models that describe and distinguish data classes for the purpose of being able to use these models to predict the class for those objects whose class label is unknown[2]. Supervised classification needs a training set to train the algorithm and test set to verify the classification accuracy that has a specific algorithm. In this work we propose the use of Random forest algorithm for some classes of stellar spectral classification of stars. We use Slogan Digital Sky Survey database to test random forest. This paper is organized as follows. Section Method introduces the random forest algorithm. Section Data describes the data used in the experiments. Section Experiments and Results shows the experimental results. Section Conclusion presents the conclusion of this work.

II. METHOD

Random forest is a recently proposed ensemble method [7] 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[8]. There are several ways to calculate output of random forest. The simplest is simple majority voting method for classification, while average output of trees regression.

OOB Error

In random forest algorithm, sampling method from train data is based on early bagging method [9] which uses bootstrap sampling method to generate different training sets. Because of randomness, nearly 37 percent of the sample will not be chosen to construct classifiers. These nearly 37 percent data are called out of bag data (OOB). These OOB data can be used to estimate the generalization error of the trees. For each tree, we get an OOB error estimate. The generalization error of random forests can be obtained by averaging all the OOB error estimates of trees.

Random Forest Algorithm

1. Draw n tree bootstrap samples from the original dataset.
2. We grow each of the bootstrap samples and gives an unpurned classification based on the following modification: at each node, we randomly sample m to try of the predictors instead of choosing the best split among those variables [10].
3. Classify new data using aggregating the results which come from the n tree (i.e., in this paper we adopt majority votes to classify dataset).

We also give an estimate of the error based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions. Calculate error rate and call it the OOB estimate of error rate.

III. 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) covering 1/3 of sky[11]. The images are taken in five photometric bands u, g, r, i and z in the optical wavelength range 0.3-1.0μm. These bands provide enough information to broadly classify these objects as stars. From the available spectra of the individual object redshift, velocity, intensity of light, temperature are calculated. As the wavelength in the available spectra ranges from 3800 to 9200 Å[11]. We choose only the visible part of wavelength from 3800 to 9200 Å.

IV. EXPERIMENTS AND RESULTS

We designed random forest classifier using Weka Software[12]. Morgan-Keenan(MK) system is widely used in astronomy classification. According to MK classification, stellar spectra were divided into totally 10 types :O, B, A, F, G, K, M, R, S, N. We generate class type as the target output of the random forest. As the wavelength in the available spectra ranges from 3800 to 9200Å ,our classifier can only classify A, F, G, K, M class types. This section shows the prediction results of the developed model. Random forest is generated using 25 trees, each tree is constructed while considering 3 random features and max depth of trees is taken to be 3.

For this the parameters set in Weka are numTrees=25, numFeatures=3 and maxDepth=3.

Table I shows the classification results using Random forest. From table 1 it is observed that this algorithm gets a good performance of about 96.60% and true positive rate refer to 100%,100%,0%,100%,100% false positive rate is 0%, 3%, 0%, 2%, 0% of Class A, F, G, K, M respectively. For five classes, we obtained a higher ROC Area 1,1,0.750,1,1.

Table II gives confusion matrix of SDSS Data using random forest algorithm, we correctly predict 6 as class A, 30 as class F,14 as class K and 7 as class M sample data. But we also wrongly predicted 1 as class F and 1 as class K sample data.

TABLE I

Performance of SDSS Data

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	A
1.000	0.034	0.968	1.000	0.984	0.967	1.000	1.000	F
0.000	0.000	0.000	0.000	0.000	0.000	0.982	0.750	G
1.000	0.022	0.933	1.000	0.966	0.955	1.000	1.000	K
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	M
0.966	0.023	0.934	0.966	0.950	0.939	0.999	0.992	Weighted Avg

Table II
Confusion Matrix of SDSS Data

a	b	c	d	e	classified as
6	0	0	0	0	a = A
0	30	0	0	0	b = F
0	1	0	1	0	c = G
0	0	0	14	0	d = K
0	0	0	0	7	e = M

To achieve optimal performance of random forest, we varied parameter numTrees and numFeatures and kept the depth of tree to be 3. numTrees was varied in {10,15,20,25,30,35,40,45,50} and numFeatures was set to 3. Fig 1 shows the relationship between Root Mean Square Error (RMSE) and different numTrees values . For each numTrees we run the program four times and got an average RMSE. From Fig. 1 we find that RMSE is lower when numtrees is 20.

We varied numFeatures in the same way. The range of numFeatures is {2,3,4,5,6,7,8,9,10} while keeping numTrees to the best 20. We got Fig. 2 which shows the relationship between numFeatures and RMSE. From Fig. 2 we can see, when the number of features are more the RMSE tends to lower. Table 3 shows the details of parameter tuning.

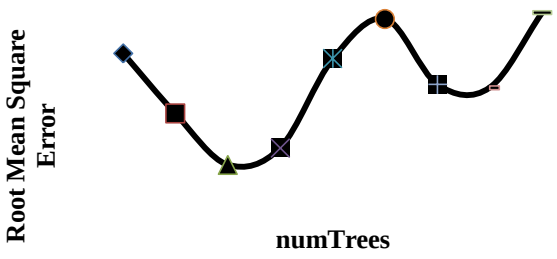

Figure 1. Relationship between numTrees and RMSE

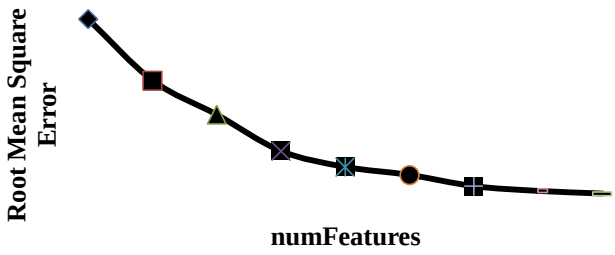

Figure 2. Relationship between numFeatures and RMSE

Table III shows that model designed gives the best performance when numTrees=20 and numFeatures=10. The corresponding RMSE is about 0.0197 and performance is about 100%.

The obtained results are compared against those of decision table, bayesian network, multilayer perceptron, random decision tree.

TABLE III
PERFORMANCE OF RANDOM FOREST WITH VARYING PARAMETERS

numtrees	numFeatures	RMSE	performance(%)
10	3	0.1299	96.6102
15	3	0.1241	96.6102
20	3	0.1192	96.6102
25	3	0.1208	96.6102
30	3	0.1294	96.6102
35	3	0.1332	96.6102
40	3	0.1269	96.6102
45	3	0.1266	96.6102
50	3	0.1338	96.6102
20	2	0.1734	91.5254
20	3	0.1192	96.6102
20	4	0.0893	98.3051
20	5	0.0572	100
20	6	0.0431	100
20	7	0.036	100
20	8	0.0261	100
20	9	0.0222	100
20	10	0.0197	100

TABLE IV
COMPARISON PERFORMANCE OF BAYESIAN, DECISION TABLE, MULTILAYER PERCEPTRON, RANDOM DECISION TREE, RANDOM FOREST

Method	Classification Accuracy(%)
Bayesian	84.7458
DT	98.3051
Multilayer Perceptron	94.9153
RDT	99.0202
Random Forest	100

Table IV shows the comparison performance of five methods. From this table we see that Random Forest get the best performance about 100 % and Bayesian obtained lowest performance of about 84.7458 %.

V. CONCLUSION

In this paper, we applied mining techniques to spectral classification. We compared the performance of different mining techniques. Our results show that ensemble learning is a better method than individual classifier. That is Random Forest Method is an effective method which is used for classification. Through tuning the parameter, we got the best performance of random forest. As future work, we are going to use more attributes that describe the astronomical objects, more kinds of these objects and large volume of data.

References

1. Kirk D. Borne. Data Mining in Astronomical Databases. Mining the sky, ESO ASTROPHYSICS SYMPOSIA. Page 671-673, 2001.
2. J. Han and M. Kamber. Data Mining, Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
3. Nong Ye. The Handbook of Data Mining. Lawrence Erlbaum Publishers, 2003.
4. Marios Kampakolou, Roberto Trotta, and Joseph Silk. Monolithic or hierarchical star formation? A new statistical analysis. Monthly Notices of the Royal Astronomical Society, 384(4):1414-1426, 2008.
5. Shaukat N. Goderya and Shawn M. Lolling. Morphological classification of galaxies using computer vision and artificial neural network: A computational scheme. Astrophysics and Space Science, 279(4):pp. 377-387, 2005.
6. Yogesh Wadadekar. Estimating photometric redshifts using support vector machines. Publications of the Astronomical Society of the Pacific, 117(827):pp. 79-85, 2005.
7. Leo Breiman, Random Forests, Machine Learning,45,5-32,2001.
8. Liaw, A.; Wiener, M. Classification and regression by randomForest, R News, 2:18-22,2002.
9. Breiman, L. Bagging predictors. Machine Learning 26,2 (1996),123-140.
10. Vladimir Svetnik; Andy Liaw; Christopher Tong.; J. Chritopher Cilberson,; Robert P. Sheridan,; Bradley P. Feuston . Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, J. Chem. Inf. Comput. Sci. [J],2003,43,1947-1958.
11. D.G. York, et al., and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. AJ,120:1579-1587, September 2000.
12. 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.

