PROCEEDINGS OF SCIENCE

A Novel Framework for Gamma-ray Source Classification using Automatic Feature Selection

Alex, P. Leung
Faculty of Information Technology, Macau University of Science And Technology
E-mail: pleung@must.edu.mo

Yuze, Tong
Faculty of Information Technology, Macau University of Science And Technology

Rui, Li
Faculty of Information Technology, Macau University of Science And Technology

Shengda, Luo
Faculty of Information Technology, Macau University of Science And Technology
E-mail: 1709853goo300061@student.must.edu.mo

C. Y. Hui[*][†]
Department of Astronomy & Space Sciences, Chungnam National University, Daejeon 34134, South Korea
E-mail: cyhui@cnu.ac.kr

With fast growing data collected by the Fermi Large Area Telescope as a big data problem, manual classification has become an impossible task for astronomers. In this paper, we propose a novel framework using machine learning techniques for gamma-ray object classification. We use the random forest (RF) algorithm for feature selection in order to achieve better classification performance. After an extensive experimental study with feature selection incoporated, we found the best results can be obtained for both active galactic nuclei (AGN) / pulsars (PSR) classification, and young (YNG) / millisecond pulsars (MSP) classification using boosted logistic regression (LR). We automate parameter tuning rather than manual tuning used in previous works. We compare the performance based on our framework with those based on Saz Parkinson et al. (2016) [1] by using the data obtained from the 3rd Fermi Large Area Telescope Source Catalog (3FGL) [2]. In PSR/AGN classification, we achieve an accuracy of $> 98\%$. On the other hand, we attain an accuracy of $> 95\%$ in the case of YNG/MSP classification.

7th Fermi Symposium 2017
15-20 October 2017
Garmisch-Partenkirchen, Germany

[*] Speaker.
[†] CYH is supported by the National Research Foundation of Korea grant 2016R1A5A1013277

1. Introduction

The advancements of astronomical instrumentations, large-scales surveys and the policy of open data access have led us into the midst of a revolution of data science. In order to fully harness the power of the deluge of data, one has to employ the techniques of machine learning and data mining. This is envisioned as the *fourth paradigm* in astronomy [3]. Machine learning has been applied in classifying objects in different wavelengths. In this paper, we focus on the γ-ray sources classification using machine learning techniques.

A number of investigations have utilized the conventional analysis to classify γ-ray sources, which requires a prior knowledge of γ-ray source properties in different classes (e.g. [4]). However, owing to the relatively short history of γ-ray astronomy, our current understanding of different classes might be far from being complete. In addition, with new data continuously pouring in and dedicated investigations, the existing frameworks are subjected to modifications and new characteristics of various classes can be established. This implies that both efficiency and accuracy of the conventional frameworks for γ-ray sources classification are unlikely to be optimal. For automatic classification, instead of relying on a prior knowledge, one let the data "speak for themselves" and generate a classification frameworks completely based on the current data. With an appropriate algorithm, attributes and patterns of the data that might be overlooked by human investigators can be possibly highlighted. As the data volume increases monotonically, automatic algorithms definitely have advantages over the traditional approaches because the new data can be trivially incorporated in most algorithms to update the model automatically.

A number of previous frameworks using the machine learning techniques are proposed for classifying Fermi γ-ray sources (e.g. [1][5] [6]). However, since the features were selected manually in these works, it is unlikely that the power of automatic classification has been fully exploited.

In this work, by coupling the classification algorithms with the automatic machine learning algorithm, we aim to improve the prediction accuracy, provide a more cost-effective prediction model, and enhance the discovery power in data mining.

2. Feature Selection

Feature selection algorithms can be used to produce a cost-effective set of predictors. Hence, it is important for improving the accuracy in the presence of a feature space with high dimensionality. In 3FGL catalog [2], there are too many potential parameters with redundancy relatively high. Methods for feature selection can produce a set of parameters with low redundancy, which leads to the more efficient training for classifiers to achieve higher accuracy.

Random Forest (RF) is an ensemble algorithm based on bagging (e.g. [7][8]). There are two significant features using RF: relatively high stability and relatively high performance. RF is essentially a tree bagging based on a Bootstrap extension. It is stable because it aggregates the predictions based on a large number of decision trees. An RF consists of many classification trees. When an object is predicted by an RF, the result of the RF takes into account the result of each classification tree with a weight.

RF has been widely used for classification in astronomy (e.g. [1][9]). Since RFs generally improve the performance of the classifiers and reduce overfitting of standard decision trees in the training set, we adopt it as the feature selection algorithm in our work.

3. The Method

We propose a novel framework to classify γ-ray sources automatically. Details of our methodology are given in [10]. There are three main stages in our framework shown in Figure 1 which illustrates (i) the preprocessing stage, (ii) the feature selection stage and (iii) the classification stage. In our framework, RFs are used for feature selection without prior knowledge after the input data is cleaned and pre-processed in the same way as [1]. In the classification stage of our framework, boosted logistic regression (LR) with the features automatically selected in stage (ii) is used to build the prediction model. The number of iteration in boosted logistic regression greatly effects the accuracy of the prediction model. This prediction model is optimized within a grid for the optimal number of iteration for the best accuracy.

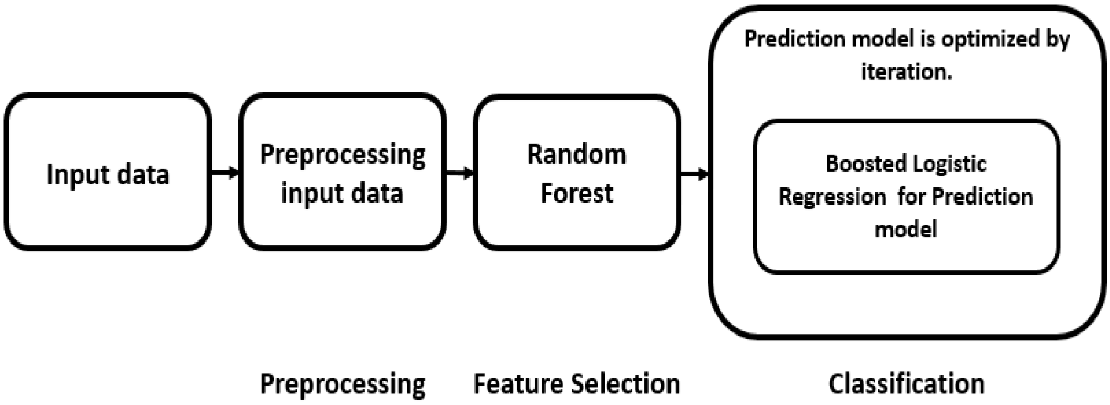
Input data
Preprocessing input data
Random Forest
Prediction model is optimized by iteration.
Boosted Logistic Regression for Prediction model
Preprocessing
Feature Selection
Classification

Figure 1: Our framework for gamma-ray sources classification (Training stage).

4. Experiments & Preliminary Results

We evaluate our framework with data from the 3FGL Catalog [2] and compare our results with those based on the method in [1]. Our experiments are divided into two parts: the PSR/AGN classification and the YNG/MSP classification. The data for PSR/AGN classification consists of 1904 γ-ray sources, while the data for YNG/MSP classification consists of 155 γ-ray sources. In each classification, the data is randomly divided into the training set (70% images) and the test set (30% images). In each part of experiments, cross-validation with 50 repeated experiments to obtain better statistics is used in our classification stage to make sure the statistical estimates for classification accuracy are reliable.

The feature automatically selected in our framework for the PSR/AGN and YNG/MSP classifications are shown in Table 1. In Figure 2, the left panel shows the comparison of receiver operating characteristic (ROC) curves for PSR/AGN classification using our model (red) and the method proposed in [1] (black) with LR as the classifier. In addition, the right panel of Figure 2 shows the comparison of ROC curves for the YNG/MSP classification. ROC curve is a plot illustrating the diagnostic ability of a binary classifier system with a discrimination threshold. The sensitivity of the ROC curve is the probability of detection, while the specific of the curve is probability of false alarm. The tradeoff between sensitivity and specificity is showed in the ROC curve which the area under is a measure of prediction model [11]. The area under the ROC curve (AUCs) and best thresholds of ROC curves produced by our framework suggest an improved performance over results obtained by [1].

In Table 2, the accuracy comparison between our framework and the approach in [1] are presented. With the use of RF for feature selection, experiment results show that our framework achieves a much higher accuracy in both types of classification over results in [1]. In PSR/AGN classification, our approach achieves an accuracy of $> 98\%$. Also, it achieves an accuracy of $> 95\%$ in the case of YNG/MSP classification.

Apart from improving the performance of supevised classification, the automatic feature selection can also provide guidelines of compiling catalogs for the future missions (e.g. Cherenkov Telescope Array). With this algorithm, one can construct a catalog with an optimal set of features for efficiently discriminating the nature of different sources and minimizing the redundancy.

Importance Rank	PSR/AGN	YNG/MSP
1	Variability_Index	Unc_Energy_Flux100
2	Signif_Curve	GLAT
3	Spectral_Index	Flux_Density
4	hr45	Signif_Curve
5	Unc_Flux1000	hr34
6	SED1000_3000	hr23
7	Flux1000_3000	Spectral_Index
8	hr23	hr45
9	Unc_Energy_Flux100	-

Table 1: Optimal sets of features for PSR/AGN and YNG/MSP classifications which are automatically selected by RF without a prior knowledge and ranked accordingly. The feature nomenclature is identical to that adopted in [1].

	Accuracy	
Prediction Model	PSR/AGN classification	YNG/MSP classification
Our method	98.2%	95.7%
Saz Parkinson et al. (2016) [1]	94.9%	90.7%

Table 2: Comparison of accuracy for both types of classification between our framework and [1] using the LR classifier.

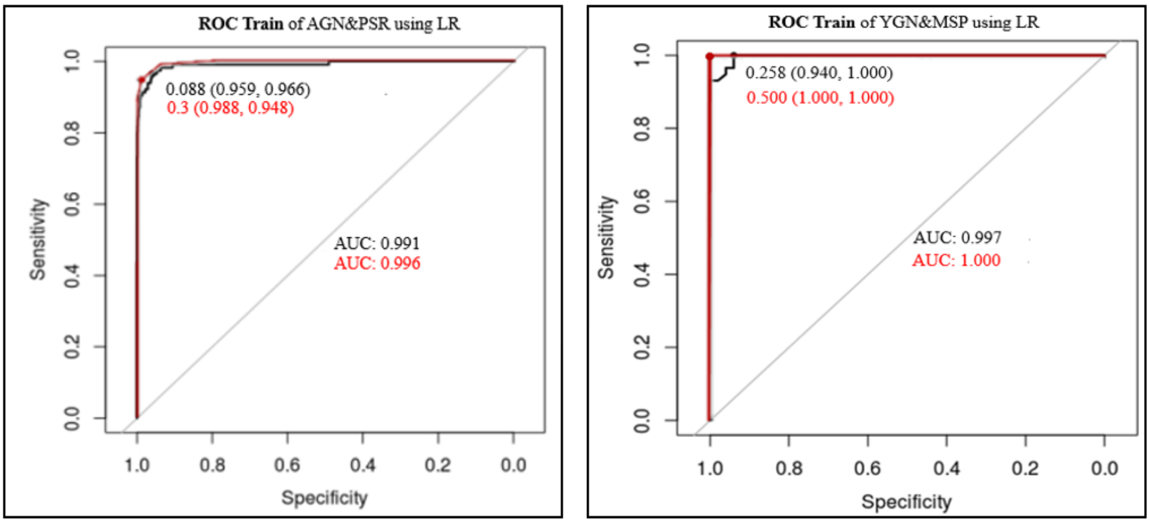

Figure 2: Comparison of ROC curves for the classification of AGN/PSR (*left panel*) and YNG/MSP (*right slide*) between our framework and [1] using the LR classifier.

References

[1] P.M. Saz Parkinson, et al. *Classification and ranking of Fermi LAT gamma-ray sources from the 3FGL catalog using machine learning techniques* *ApJ, 820, 8* (2016) [arXiv:1602.00385].

[2] F. Acero, et al. *Fermi large area telescope third source catalog* *ApJS, 218, 23* (2015) [arXiv:1501.02003].

[3] G. Bell, T. Hey, & A. Szalay, *Beyond the Data Deluge* *Science, 323, 1297* (2009).

[4] C.Y. Hui, et al. *Searches for Millisecond Pulsar Candidates among the Unidentified Fermi Objects* *ApJ, 809, 68* (2015) [arXiv:1507.02604].

[5] M. Ackermann, et al. *A Statistical Approach to Recognizing Source Classes for Unassociated Sources in the First Fermi-LAT Catalog* *ApJ, 753, 83* (2012) [arXiv:1108.1202].

[6] N. Mirabal, et al. *Fermi's SIBYL: mining the gamma-ray sky for dark matter subhaloes* *MNRAS, 424, L64* (2012) [arXiv:1205.4825].

[7] Tin Kam Ho *Random Decision Forests* *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 August 1995. pp. 278-282* (1995).

[8] Breiman, Leo *Random forests* *Machine Learning, vlo. 45, no. 1, pp. 5-32* (2001).

[9] D. Carrasco et al. *Photometric classification of quasars from RCS-2 using Random Forest* *A&A, 584, A44* (2015).

[10] Alex, P. Leung et al. *Application of automatic feature selection algorithms in classifying gamma-ray objects* submitted to ApJ (2017).

[11] Fawcett, Tom *An introduction to ROC analysis* *Pattern recognition letters, vlo. 27, no. 8, pp. 861-874* (2006).