Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 1096–1101

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Statistical Methods and Artificial Intelligence
(IWSMAI 2020)
April 6-9, 2020, Warsaw, Poland

Detection of Pulsar Candidates using Bagging Method

Mourad Azhari[a,*], Abdallah Abarda[b], Altaf Alaoui[a], Badia Ettaki[c], Jamal Zerouaoui[a]

[a]*Laboratory of Engineering Sciences and Modeling, Faculty of Sciences- Ibn Tofail University, Campus Universitaire, BP 133, Kenitra, Morocco*
[b]*Laboratoire de Modélisation Mathématiques et de Calculs Economiques, FSJES, Université Hassan 1er, Settat, Morocco*
[c]*Laboratory of Research in Computer Science, Data Sciences and Knowledge Engineering, Department of Data, Content and knowledge Engineering School of Information Sciences Rabat, Morocco*

Abstract

The pulsar classification represents a major issue in the astrophysical area. The Bagging Algorithm is an ensemble method widely used to improve the performance of classification algorithms, especially in the case of pulsar search. In this way, our paper tries to prove how the Bagging Method can improve the performance of pulsar candidate detection in connection with four basic classifiers: Core Vector Machines (CVM), the K-Nearest-Neighbors (KNN), the Artificial Neural Network (ANN), and Cart Decision Tree (CDT). The Error Rate, Area Under the Curve (AUC), and Computation Time (CT) are measured to compare the performance of different classifiers. The High Time Resolution Universe (HTRU2) dataset, collected from the UCI Machine Learning Repository, is used in the experimentation phase.

Keywords: Bagging; CVM; ANN; KNN; CDT; AUC; Error Rate; Computation Time(CT)

1. Introduction

The detection of pulsar candidates constitutes a great challenge for scientists in the astrophysics area. Machine learning is proposed to solve the problem of the pulsar signal. Many works relevant to pulsar candidate detection have been advanced in the astrophysics field. Eatough et al. [1] applied the Artificial neural network algorithm to identify the pulsars as a new method exceeding the maximum likelihood method. Bates et al.[2] used a basic statistic, obtained from the pulsar candidate plots, as inputs to the Artificial Neural Network. Loginov and Malov [3] used Principal Component Analysis (PCA) to split short-period and long-period pulsars into two classes. Similarly, Lee et al. [4] advanced the pulsar evaluation algorithm for candidate extraction (PEACE) as a new alternative for visual inspection

* Corresponding author. Tel.: +212694324310
E-mail address: azharimourad@yahoo.fr

1877-0509
10.1016/j.procs.2020.03.062

and pulsar candidate detection. This method is efficient to identify pulsar signals. Lyon et al. [5] proposed the Gaussian Helling Very Fast Decision method (GHVFDM) as an option of manual detection of pulsars. GHVFDM is applied to Big Data, generated by Square Kilometer Array (SKA) which is capable of handling millions of candidates at second. In this work, we use the Bagging Method combined to different classifiers as Core Vector Machines (CVM), K-nearest neighbors (KNN), Artificial Neural Network (ANN), and Cart Decision Tree (CDT) to evaluate the impact of the combined technique, on a HTRU2 pulsar dataset by using prediction power(Error Rate), Area under the Curve (AUC) and Computation Time (CT) metrics. The following paper will tackle the major related works in the first section. The second section will propose machine learning methods to detect pulsar candidates. While the fourth section will describe the HTRU2 dataset. The fifth section will present our experimental results and their analysis.

2. Related work in Bagging Method

The performance of the classification models, such as SVM, KNN, ANN and CDT and so on [6, 7], has been improved in several studies through the implementation of the Bagging Technique. Indeed, Breiman [8] employed the Bagging predictors to generate different forms of a predictor; he used various tests on real and simulated data sets, applied in the classification and regression tree. Zheng [9] tested the Boosting and Bagging Algorithms with the neural network on time series in the finance area as the classification performance of these two methods exceed the SVM method and the variance of logistic regression in the field of financial management. However, Alfaro et al. [10] compared the accuracy metric of two AdaBoost algorithms which are the AdaBoost with Decision Tree and artificial neural networks. The first algorithm works more efficiently than the basic ANN algorithm. In the same way, Mordelet and Vert [11] suggested an application of the Bagging (SVM) in " the Pattern Recognition Letters " data to solve the learning problem from positive and unlabeled instances. The performance of this model is well in comparison to current approaches. Kim et al. [12] obtained the same results using multiple simulations of " Iris" and "hand-written digit recognition" datasets. Furthermore, Ford has used many machine learning methods to solve the problem of pulsar candidates. [13, 14, 15, 16].

3. Proposed methods

In the following paper, we have applied different algorithms to this pulsar classification problem: CVM, KNN, ANN, CDT, and Bagging Method.

3.1. Core Vector Machine (CVM) Method

Core vector Machine (CVM) is a technique for scaling up two classes which give similar results of the SVM method in the case of Big Data sets manipulation. In CVM, the quadratic optimization problem implicated in SVM is conceived as an equivalent to the Minimum Enclosing Ball (MEB) problem [17]. At the output, like the SVM, this technique produces a series of support points that can be used for prediction.

3.2. K-Nearest Neighbors (KNN) Method

K-Nearest Neighbors (KNN) algorithm is based on two steps, the first is to memorize all available instances and the second is to classify the news instances using a similarity measure. We generally look at two important aspects: Ease to interpret the output and Predictive Power [18].

3.3. Artificial Neural Network (ANN) Method

Artificial Neural Network (ANN) method is a family of machine learning algorithms inspired by the biological system of interconnected neurons [19] to classify and treat mapping problems. ANN method aims is to learn the non-linear function and map the input features to output classes.

3.4. Cart Decision Tree (CDT) Method

Cart Decision Tree method (CDT) [20] is an algorithm used to create a training model to predict classes or values of the target variables via deduction of the learning decision rules from the training data. CDT presents a simple visualization of results and it is a basic predictor for the Bagging Model [21].

3.5. Bagging Method

Bagging is an ensemble method conceived to improve the stability and accuracy of machine learning algorithms adapted to classification and regression. It is capable to reduce the variance, resists to over-fitting, added to its ability to be used with any basic Algorithm [8]. Bagging Algorithm works in two stages:

- Training stage: In each iteration t, t=1,..T:

1. Randomly sample with replacement N samples from the training set;
2. Train a "selected base classifier" (CVM, KNN, ANN, CDT) on the samples.

- Test stage: For each test example:

1. Start all trained base models;
2. Predict by combining results;
3. for all T trained Classification model: output = "majority vote".

4. Distribution of HTRU2 dataset

The following paper uses the HTRU2 dataset downloaded from UCI site [12]. The class label (signal, background event) is a dependent response, wichi is followed by eight pulsar features as independent variables:

- Statistics of the folded pulse profile: Mean, Standard deviation, Excess Kurtosis, and Skewness of the integrated profile[13].
- Statistics based on the DM-SNR curve: Mean, Standard deviation, Excess Kurtosis, and Skewness of the DM-SNR curve [13].

The class label consists of 9% of the signal and 91% of background events. Hence, We conclude that the distribution of the HTRU2 dataset is imbalanced.Consequently, handling imbalanced dataset requires the use of the resampling techniques as the data split and the K-Folds Cross-validation.

4.1. Data Split

Looking to get an unbiased estimation of the performances of the algorithms, the HTRU2 dataset is divided into a training sample (70%) and test sample (30%). The distribution of the target variable in the test-sample is confirmed with the initial distribution. Hence, we can confirm the representativeness of the test sample (see Table 1).

Table 1. Distribution of the class label in the initial and test sample

Initial sample distribution		Test-sample distribution	
background	signal	background	signal
0.90842552	0.09157448	0.90883669	0.09116331

4.2. K-Folds cross-validation

K-Folds cross-validation is a technique that involves reserving a sample of a dataset [22]. In this way, the steps of cross-validation are:

- Randomly split the entire dataset into k-folds;
- For each k-fold in the dataset, build the model on K-1 folds of the dataset, testing the model for the reprsentativeness of K th folds;
- Record the error observed on each of the predictions;
- Repeat this until each of the k-folds has served as the test set.

In the practice and as a general rule, we choose k=10 [23].

5. Experimental results and analysis

5.1. Evaluation function

We present the important metrics to evaluate the classifier performance: Accuracy, AUC-ROC, and Computation Time (CT).

5.1.1. Accuracy metric

The equations below present some of the important metrics that are evaluated for comparison [24].

- False Positive Rate (FPR) (Type I Error): Number of examples wrongly classified as the signal out of total true non-pulsar signal:

$$FRP = \frac{FP}{FP + TN} \tag{1}$$

- False Negative Rate (FNR) (Type II Error): Number of instances wrongly classified as non-pulsar signal out of the total true presence of pulsar signal:

$$FNR = \frac{FN}{FN + TP} \tag{2}$$

- Accuracy is the proportion of total examples classified correctly:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 1 - ErrorRate \tag{3}$$

Where:

- TP: The true positive is the number of candidate examples that are pulsars and are being classified as pulsars;
- TN: The true negative is the number of candidates which are non-pulsars and being classified as non -pulsars;
- FN: The false negative is the number of true pulsar candidates that are wrongly being classified as non- pulsars;
- FP: The false positive is the number of non-pulsars candidates that are wrongly being classified as pulsars.

5.1.2. ROC-AUC metric

Receiver Operating Characteristic (ROC) Curve allows comparing various supervised learning classifiers. It is especially useful for cases of skewed class distribution [25]. Area Under Curve (AUC) is an area equivalent to the probability that the algorithm will place a randomly selected positive example higher than a randomly selected negative example [26].

5.1.3. Computation time (CT) metric

The Computation time estimates the running time required to perform a computational process. In this paper, all Algorithms were trained using a machine with Intel (R) cores (TM) i7,7500U CPU@ 2.7 GHz 2.9 GHz, 8 Go memory (RAM) and processor x64.

5.2. Experimental results and analysis

5.2.1. Experimental results

Table 2 presents the metric evaluations: Test Set Error Rate, 10-Folds cross-validation, and their consequent variations, AUC and CT.

Table 2. The metric evaluations:Test Set Error Rate, 10-Folds, Variation, AUC and CT

	Test Set Error Rate	10-Folds Cross Validation	Variation	AUC	CT(s)
CVM	0,0199	0,0214	0,0015	0,5715	5,29
KNN	0,0227	0,0217	0,001	0,994	32,03
ANN	0,029	0,0207	0,0083	0,9743	2,83
CDT	0,0231	0,0212	0,0019	0,9249	0,19
Baggin(CVM)	0,0218	0,0215	0,0003	0,9377	333,34
Baggin(KNN)	0,0199	0,0213	1,00E-04	0,9913	2398,2
Bagging(ANN)	0,0207	0,0206	1,00E-04	0,9538	11,47
Bagging(CDT)	0,0218	0,0211	0,0007	0,9775	2,89

5.2.2. Analysis

We compare different models on HTRU2 dataset in terms of AUC, Error Rate, and computation time metrics.

1. Comparison of basic classifiers: CVM, KNN, ANN, and CDT
   - The measure of the Error Rate proves that all basic classifiers work efficiently. The predictive power varies between 1.99% and 2.31%. This conclusion is confirmed with the 10-Folds cross-validation method (see figure 1).
   - The measure of the AUC shows that the KNN model is more powerful (0.994) than the ANN (0.9743) and CDT (0.9249) models. While the CVM model records a low performance (0.5715) (see figure 1).
   - The computation time measure shows that the running time classifier is very fast for CDT (0.19 s) and ANN (2.83 s), fast for CVM (5.29 s) and slow for KNN (32.03 s).
2. Comparison of Bagging with different classifiers: CVM, KNN, ANN, and CDT.
   - In terms of the error rate metric, the predictive power of Bagging Methods is well-performant. It is nearly close to the results, recorded by the basic models (between 2.07% and 2.18 %). Also, this conclusion is confirmed by the 10-Folds Cross-Validation Method (see Table 2).
   - Concerning the AUC measure, the Bagging Method connected to basic classifiers runs efficiently. The performance varies between 92% at Bagging (CVM) and (99%) at Bagging (KNN) that is better than Bagging (CDT) with (97.75%) and Bagging (ANN) with 95.38%. Hence, we conclude that Bagging impacts positively CVM (+36%) and CDT (+5%). However, the KNN and ANN classifiers are closes to stability (99%) and any improvement is noted.
   - The computation time metric proves that the Bagging (CDT) works more rapidly (2.89s) than the Bagging(ANN) (11.47s). While the Bagging (CVM) reacts slowly (333.34s) and the Bagging(KNN) works very slowly (2398.17s). Figure 1 shows the lost time produced, using the Bagging method. Then, It clear that the classifiers are penalized, especially KNN (2366.17 s) and CVM (328.05 s) (see figure 1).

6. Conclusion

In the context of HTRU2 dataset, as an imbalanced distribution, we conclude that:

- minimal improvement is noted in term of test set error rate;
- the Bagging Method affects positively the CVM classifier in term of AUC measure.
- the Bagging Algorithm penalizes KNN and CVM models in term of computation time;
- the Bagging results are similar to 10-Folds Cross-Validation;
- Bagging runs efficiently to adjusting the dataset distribution.

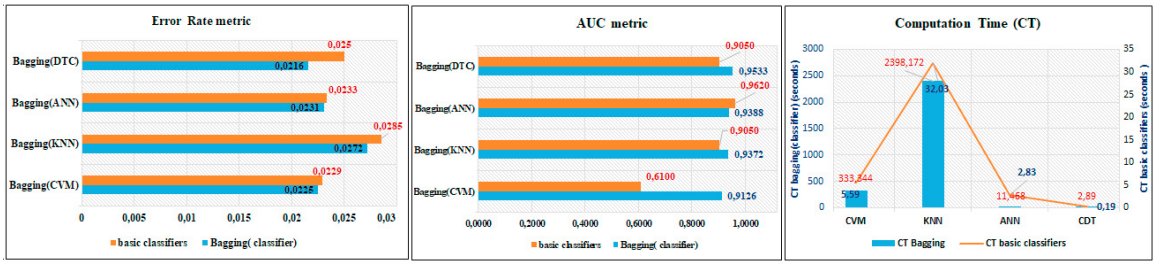

Fig. 1. Error rate metric , AUC metric, and computation time of Baggin(classifiers)

References

[1] Eatough, R.P et al. (2010) "Selection of radio pulsar candidates using artificial neural networks: Selection of radio pulsar candidates", *Mon. Not. R. Astron. Soc* **407**: 2443–2450.

[2] Bates, S et al.(2012) "The High Time Resolution Universe Pulsar Survey – VI. An artificial neural network and timing of 75 pulsars: HTRU – VI. An ANN and 75 normal pulsars, *Mon. Not. R. Astron. Soc* **427**: 1052–1065.

[3] Loginov, A., and Malov, I. (2013)"Classification of pulsars using the principle-components method"*Astron. Rep* **57**:1001–1013.

[4] Lee, K et al. (2013)"peace: pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates"*Mon. Not. R. Astron. Soc* **433**: 688–694.

[5] Lyon,R, Steppers, B., Cooper,S. , Brooke, J., and Knowles,J. (2016)" Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach" *Mon. Not. R. Astron. Soc* **459**: 1104–1123.

[6] Abarda, A., Bentaleb, Y.,and Mharzi, H. (2017)" A divided latent class analysis for big data" *Procedia Computer Science* **110**: 428-433.

[7] Abarda, A., Bentaleb, Y., El Moudden, M., Dakkon, M., Azhari, M., Zerouaoui, J., and Ettaki, B.(2018)"Solving the problem of latent class selection"*In Proceedings of the International Conference on Learning and Optimization,ACM Algorithms: Theory and Applications* **15**

[8] Breiman, L. (1996) "Bagging predictors *Mach. Learn*, **24**: 123–140.

[9] Zheng, Z.(2006) "Boosting and Bagging of Neural Networks with Applications to Financial Time Series" .

[10] Alfaro,E., García,N., Gámez, M., and Elizondo, D.(2008)"Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks" *Support Syst* **45**: 110–122.

[11] Mordelet, F., and Vert,J.-P. (2014) "A Bagging SVM to Learn from Positive and Unlabeled Examples." *Pattern Recognition Letters* **37**: 201—209. https://doi.org/10.1016/j.patec.2013.06.010.

[12] Kim, H., Pang, S. Je,H., Kim, D., and Bang S.(2002)"Support Vector Machine Ensemble with Bagging" *in Pattern Recognition with Support Vector Machines* **2388**: 397–408.

[13] Ford, J.(2017) "Pulsar Search Using Supervised Machine Learning, Doctoral dissertation" *Nova SoutheasternUniversity,Retrieved from Nsuw-Works, College of Engineering and Computing*

[14] Azhari, M., Alaoui, A., Abarda, A., Ettaki, B.,and Zerouaoui , J. (2020) "Using Ensemble Methods to Solve the Problem of Pulsar Search" ıIn: Farhaoui Y, (eds) Big Data and Networks Technologies, BDNT 2019, Lecture Notes in Networks and Systemst,Springer **81**: 183-189.

[15] Azhari, M., Alaoui, A., Achraoui Z(2019)."Ettaki, B.,and Zerouaoui, J.: Adaptation of the Random Forest Method: Solving the problem of Pulsar Search"*The Fourth International Conference on Smart City Applications,ACM* doi: 10.1145/3368756.3369004

[16] Azhari, M., Alaoui, A., Abarda A., Ettaki, B.,and Zerouaoui, J.(2020): "A Comparison of Random Forest Methods for Solving the Problem of Pulsar Search" *The Fourth International Conference on Smart City Applications, Springer, Cham.*

[17] Tsang,I., Kwok,J., and Cheung, P.(2015) "Core Vector Machines Fast SVM Training on Very Large Data Sets" *Journal of Machine Learning Research* **6**: 363—392.

[18] Altman, N.(1992)"An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression" *Am. Stat* **46**: 175–185.

[19] Dawson,C., and Wilby, R.(1998)"An artificial neural network approach to rainfall-runoff modelling" *Hydrol.Sci* **43**:47–66.

[20] Loh,W.(2011)"Classification and regression trees" *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**: 14–23.

[21] Genuer,R., and Poggi, J.(2017)" Arbres CART et forets al éatoires-Importance et sélection de variables" .

[22] Browne, M.W.(2000) "Cross-Validation Methods" *Journal of Mathematical Psychology* **44**: 108–132 https://doi.org/10.1006/jmps.1999.1279

[23] Anguita, D, Ghelardoni, L, Ghio, A, Oneto, L, and Ridella, S(2012) "The K in K-fold Cross Validation" *proceedings European Symposium on Artificial Neural Networks, Computational Intelligence, and Machine Learning, Bruges (Belgium)*: 441–446

[24] Fawcett,T.(2006)"An introduction to ROC analysis" *Pattern Recognit* **27**:861–874.

[25] Kumar, R., and Indrayan, A.(2011) "Receiver operating characteristic (ROC) curve for medical researchers"*Indian Pediatrics* **48**: 277—287.

[26] Lobo,J. M.,Jiménez-Valverde, A., and Real,R.(2008) "AUC: a misleading measure of the performance of predictive distribution models" *Global Ecology and Biogeography* **17**: 145–151.