Weighting Features Before Applying Machine Learning Methods to Pulsar Search

Dayang Wang, Qian Yin, Hongfeng Wang

Weighting Features Before Applying Machine Learning Methods to Pulsar Search

Dayang Wang, Qian Yin* and Hongfeng Wang

Image Processing and Pattern Recognition Laboratory, Beijing Normal University, Beijing 100875, China
wdyan9@163.com, *yinqian@bnu.edu.cn, dzuwhf@163.com

Abstract. In recent years, different Artificial Intelligence methods have been applied to pulsar search, such as Artificial Neural Network method, PEACE Sorting Algorithm, Real-time Classification method. In this paper, Weighting Feature method before applying machine learning (ML) was proposed. We give weight to each feature according to its ability to distinguish pulsar and non-pulsar candidates. The ability is determined by the separation degree of the distribution of pulsars and non-pulsars on particular feature. And then use the ML methods to classify different types of candidates. The results show that this method is significant. The accuracy of identifying pulsars and modeling time were both improved after weighting.

Keywords: Weighting, Machine Learning, Pulsar Search, WEKA.

1 Introduction

Pulsar is fast rotated neutron star, which periodically sends pulse signal whose period is short and very stable. Pulsar plays an important part in physics, astronomy and many other fields. In recent years, AI methods like image pattern recognition [2], artificial neural network method and scheduling algorithm are used in pulsar search. Lee et al. (2013) proposed the PEACE sorting algorithm to search pulsar, which had obtained good results. Lyon et al. (2016) used the GH-VFDT (Gauss-Hellinger Very Fast Decision Tree) to distinguish the candidate, with recognition rate of pulsars over 90% [3].

While GH-VFDT obtained a high recognition rate of pulsars, the difference between the abilities of different features to distinguish the pulse and non-pulsar are not reflected. Thus, in this paper, we add different weights to the eight features before the machine learning process according to their separation degree. Results show that weighting improves both the accuracy rate of classification and modeling time.

The structure of this paper is shown as follows: the related work is mentioned in section 2; the Feature Weighting method is proposed in section 3; and with its corresponding experiments are showed and analyzed in section 4 and 5; the section of conclusion comes as the end.

2 Related Work

2.1 Feature

In the process of searching for pulsar signals with radio telescope, the most basic data are obtained. These data are subjected to Removing signal interference, de dispersion, FFT [4] and periodic search. Then a pulsar candidate is generated which has some basic Features. Lyon et al. (2016) used eight new features to describe the pulsar candidate. The eight features are Mean of the integrated profile $Prof_{\mu}$, Standard deviation of the integrated profile $Prof_{\sigma}$, Excess kurtosis of the integrated profile $Prof_{k}$, Skewness of the integrated profile $Prof_{s}$, Mean of the DM-SNR curve DM_{μ}, Standard deviation of the DM-SNR curve DM_{σ}, Excess kurtosis of the DM-SNR curve DM_{k}, Skewness of the DM-SNR curve DM_{s} [5]. Pulsar Feature Lab and Presto [6] are used to process the primitive data into these eight features.

2.2 Dataset

Three separate datasets were used to the measure the performance of ML methods on pulsar search. The small scale dataset is LOTAAS which was obtained during the LOTAAS survey and is currently private. The medium scale dataset HTRU2 was obtained during an analysis of HTRU Medium Latitude data by Thornton (2013). The large scale dataset HTRU1 is produced by Morello et al. The detailed information of the three datasets is summarized in the table1.

Table 1. Three pulsar candidate datasets

Dataset	Creator	Time	Volume	Pulsar	Non-pulsar
LOTAAS	Morello et al	2012	5053	66	4987
HTRU2	Thornton	2013	17898	1639	16259
HTRU1	Lofar	2013	91191	1196	89995

3 Methodology−Feature Weighting

Analyzing the statistic distribution of the eight features from the sample data of pulsars and non-pulsars, feature data was extracted from 90, 000 labelled pulsar candidates produced by Morello et al. (2014), via Pulsar Feature Lab. As it is showed in figure1, the data were scaled to the interval of [0, 1]. For each feature, there are two box plots. The orange red box shows the feature distribution of known pulsars, while the box in light blue describes the RFI/noise distribution.

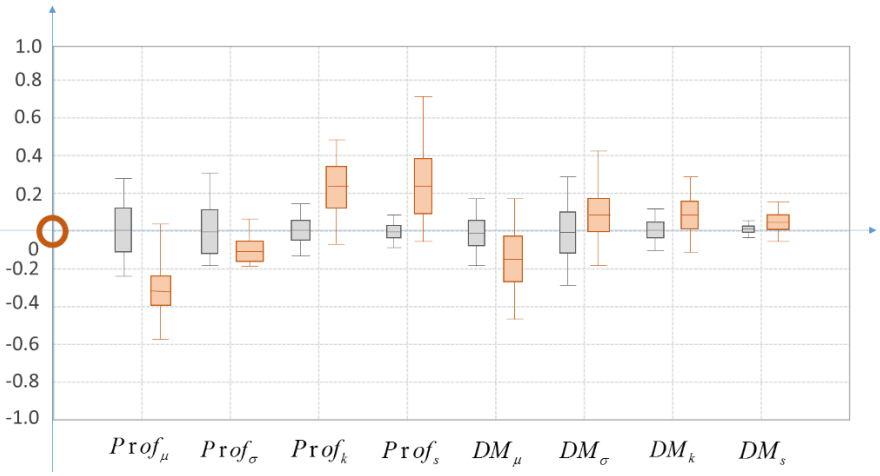

Fig. 1. Feature distribution of pulsars and non-pulsars

It is obvious that when we are classifying a pulsar candidate via its feature, the feature that has a high degree of separation between pulsars and non-pulsars weighs more than other features. Therefore, this paper naturally adds different weights to the eight features according to their separation degree between different types of candidates. As a specific feature, this paper defines the separation degree as follows:

$$Ab = -\frac{l}{R_P} - \frac{l}{R_I} + 2 \tag{1}$$

In this formula, as can be seen from figure2, for a particular feature, *Ab* denotes the separation degree, *l* means the coincident area of pulsar and non-pulsar, *Rp* denotes the width of the distribution of the pulsars on the feature, while R_I means the distribution width of non-pulsars.

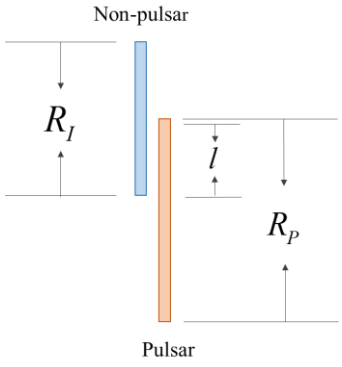

Fig. 2. Distribution of pulsar and non-pulsars and their coincident area

The distribution of features between candidates can be considered as natural distribution. According 3σ principle, features of almost all candidates will be within the range

of feature box. By analyzing the data from LOTAAS, HTRU2 and HTRU1, this paper get the weight of each feature W_i (i = 1~8).

$$W = (1.64, 0.85, 1.91, 1.38, 1.29, 1.50, 2.03, 1.18) \quad (2)$$

4 Experiments

In this part, weighting each feature before utilizing ML methods on the datasets are proposed. Classification accuracy and modelling time are both taken to be criterion to judge the performance of the methods. The paper supposes weighting is useful if methods improves the accuracy or improves the modeling time. What's more, accuracy goes before modeling time.

Table 2. Accuracy and modeling time before and after weighting for LOTAAS.

Methods	Accuracy		Modeling time	
	Before	After	Before	After
SMO	99.7625%	99.8813%	0.05	0.02
IBK	99.8417%	99.8615%	0	0
JRIP	99.8417%	99.8615%	0.08	0.11
J48	99.8615%	99.8615%	0.05	0.03
RandomForest	99.8615%	99.8615%	0.7	0.47

For small scale dataset LOTAAS, the experimental results are shown in Table2, after weighting, accuracy rate of SMO, IBK and JRIP are improved. Modeling time of J48 and RandomForest are improved.

Table 3. Accuracy and modeling time before and after weighting for HTRU2.

Methods	Accuracy		Modeling time	
	Before	After	Before	After
SMO	97.5640%	97.5696 %	0.22	0.52
IBK	97.1449 %	97.1449 %	0	0
JRIP	97.8154 %	97.8154 %	1.84	1.44
J48	97.8433 %	97.8433 %	0.28	0.2
RandomForest	97.9942 %	97.9942 %	9.94	7.16

For medium scale dataset HTRU2, the experimental results are shown in Table3, after weighting, accuracy rate of SMO is improved. Modeling time of JRIP, J48 and Random-Forest are improved, while IBK remains the same.

Table 4. Accuracy and modeling time before and after weighting for HTRU1.

Methods	Accuracy		Modeling time	
	Before	After	Before	After
SMO	99.5866 %	99.5866%	0.36	0.36
IBK	99.5340 %	99.5340%	0.06	0.01

JRIP	99.6129 %	99.6228 %	4.78	18.0
J48	99.6074 %	99.6085 %	3.16	3.23
RandomForest	99.6721 %	99.6798 %	77.72	83.8

For large scale dataset HTRU1, as is shown in table4, after weighting, accuracy rate of JRIP, J48 and RandomForest are improved. Modeling time of IBK is improved, while SMO remains the same.

In conclusion, for the five ML methods SMO, IBK, JRIP, J48 and RandomForest, weighting either improves the accuracy or modeling time, or in the worst cases, weighting will at least be the same as not weighting.

5 Discussions

This part explains why SMO, IBK, JRIP, J48 and RandomForest are selected to test the effects of weighting instead of other ML methods. In this paper, we actually experimented various ML methods using WEKA.

Table 5. Accuracy rate of pulsar recognition of various ML methods before weighting

Types	Methods	LOTAAS	HTRU2	HTRU1
bayes	NaiveBayes	99.5448%	94.4966 %	98.9155 %
func-	LibSVM	98.7928%	91.1443 %	98.8585 %
tions	SMO	99.7625%	97.5640%	99.5866 %
lazy	IBK	99.8417%	97.1449 %	99.5340 %
	LWL	99.9010%	97.7539 %	99.5175 %
meta	AdaBoostML	99.8615%	97.6534 %	99.5175 %
	Stacking	98.6938%	90.8426 %	98.6885 %
misc	InputMappedClassifier	98.6938%	90.8426 %	98.6885 %
rules	JRIP	99.8417%	97.8154 %	99.6129 %
trees	HoeffdingTree	99.8417%	97.4411 %	99.5822 %
	J48	99.8615%	97.8433 %	99.6074 %
	RandomForest	99.8615%	97.9942 %	99.6721 %
	RandomTree	99.8021%	96.8432 %	99.5241 %

In table5, the purple number means the corresponding methods performs better than others. As is shown, with the scale of datasets becomes larger, SMO, IBK, JRIP, J48 and RandomForest have better performance over other algorithms.

6 Conclusion

Due to its stable cycle, Pulsar plays a very important part in physics, astronomy and many other fields. Traditional ways of pulsar search are manual. In recent years, Artificial intelligence is widely used in various fields and achieves great success. Therefore, AI methods are gradually applied to pulsar search. This paper is based on the principled real-time classification approach. Eight features are used to describe a pulsar candidate.

Before applying ML methods on datasets, this paper weights each feature according to their separation degree, and then find out that either the accuracy or modeling time is improved after weighting.

Acknowledgments. The research work in this paper was supported by the grants from National Natural Science Foundation of China (61472043, 61375045) and the Joint Research Fund in Astronomy (U1531242) under cooperative agreement between the NSFC and CAS, Beijing Natural Science Foundation (4142030). Prof. Qian Yin is the author to whom all the correspondence should be addressed.

References

1. Zhichen Pan, Lei Qian, Youling Yue. Pulsar search technology and FAST telescope pulsar search foreground [J]. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012.2016.
2. W. W. Zhu1, A. Berndsen1, E. C.Madsen1, M. Tan1, I. H. Stairs1, A. Brazier2, P. Lazarus3, R. Lynch4, P. Scholz4,K. Stovall5,6, S. M. Ransom7, S. Banaszak8, C. M. Biwer8,9, S. Cohen5, L. P. Dartez5, J. Flanigan8, G. Lunsford5,J. G. Martinez5, A.Mata5, M. Rohr8, A. Walker8, B. Allen8,10,11, N. D. R. Bhat12,13, S. Bogdanov14, F. Camilo14,15,S. Chatterjee2, J.M. Cordes2, F. Crawford16, J. S. Deneva17, G. Desvignes3, R. D. Ferdman4,18, P. C. C. Freire3,J. W. T. Hessels19,20, F. A. Jenet5, D. L. Kaplan8, V. M. Kaspi4, B. Knispel10,11, K. J. Lee3, J. van Leeuwen19,20,A. G. Lyne18, M. A.McLaughlin21, X. Siemens8, L. G. Spitler3, and A. Venkataraman15. SEARCHING FOR PULSARS USING IMAGE PATTERN RECOGNITION. University of British Columbia.2014.
3. Yuyun Xu, Chenchen Fan. Application of Artificial Intelligence in Pulsar Screening [J]. National Astronomical Observatories, Chinese Academy of Sciences, Beijing. 2006.
4. Qingyong Zhou, Jianfeng Ji, Hongfei Ren. X-ray pulsar cycle fast search algorithm for non-equal interval time data [J]. National Key Laboratory of Geographic Information Engineering, Xi'an 710054. 2012.
5. R. J. Lyon 1⋆, B. W. Stappers 2, S. Cooper 2, J. M. Brooke 1, J. D. Knowles 1,3. Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach. The University of Manchester. 2016.
6. K. J. Lee,1‹ K. Stovall,2,3 F. A. Jenet,2 J. Martinez,2 L. P. Dartez,2 A. Mata,2G. Lunsford,2 S. Cohen,2 C. M. Biwer,4 M. Rohr,4 J. Flanigan,4 A. Walker,4S. Banaszak,4 B. Allen,4,5 E. D. Barr,1 N. D. R. Bhat,6,7 S. Bogdanov,8 A. Brazier,9F. Camilo, 8, 10 D. J. Champion, 1 S. Chatterjee,9 J. Cordes,9 F. Crawford,11 J. Deneva,10G. Desvignes, 1 R. D. Ferdman, 12, 13 P. Freire,1 J. W. T. Hessels,14,15 R. Karuppusamy,1V. M. Kaspi, 12 B. Knispel, 5 M. Kramer, 1,13 P. Lazarus,1 R. Lynch,12 A. Lyne,13M. McLaughlin, 16 S. Ransom, 17 P. Scholz, 12 X. Siemens,4 L. Spitler,1 I. Stairs,18M. Tan, 18 J. van Leeuwen14, 15 and W. W. Zhu18. PEACE: pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates. Royal Astronomy Society. 2013-5-27.
7. Meiyu Yuan. Application of Data mining and Machine learning using WEKA. 2nd edition. Tsinghua University Press. 2016.8. 498-526.
8. V. Morello1, 2, E.D. Barr1, 2, M. Bailes1,2, C.M. Flynn1, E.F. Keane1,2and W. van Straten1,2. SPINN: a straightforward machine learning solution to the pulsar candidate selection problem. Swinburne University of Technology.2014.