Improving Pulsar Candidate Identification with Grid Group Uniform Sampling Postprint

Authors: Yi-Ning Song, Mao-Zheng Chen and Zhi-Yong Liu

Date: 2025-06-13T16:51:13+00:00

Abstract

Pulsar candidate identification is an indispensable task in pulsar science. Based on the characteristics of imbalanced and diverse pulsar data sets, and the lack of a unified processing framework, we first used dimensionality reduction and visualization to analyze potential deficiencies caused by the incompleteness of current data set extraction methods. We found that the limited use of non-pulsar data may lead to bias in the result, which may limit the generalization ability. Based on the dimensionality reduction results, we propose a Grid Group Uniform Sampling (GGUS) method. This data preprocessing method improves the performance of Random Forest, Support Vector Machine, Convolutional Neural Network, and ResNet50 models on Lyon' s features, diagnostic plots, and period-dispersion measure (period-DM) plots in the HTRU1 data set. The average recall increased by approximately 0.5%, precision by nearly 2%, and F1 score by around 1.2% for all models and in all data sets. In the period-DM plots testing, the high-performance ResNet50 algorithm achieved over 98% F1 score using random sampling. GGUS demonstrated further improvements in this test, enhancing the average F1 score, precision, and recall by approximately 0.07%, 0.1%, and 0.03%, respectively.

Full Text

Improving Pulsar Candidate Identification with Grid Group Uniform Sampling

Yi-Ning Song[12], Mao-Zheng Chen[13], and Zhi-Yong Liu[13]
[1]Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China; chen@xao.ac.cn, liuzhy@xao.ac.cn
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Radio Astronomy and Technology (Chinese Academy of Sciences), Beijing 100101, China

Received 2024 September 18; revised 2025 January 10; accepted 2025 January 20; published 2025 May 8

Abstract

Pulsar candidate identification is an indispensable task in pulsar science. Based on the characteristics of imbalanced and diverse pulsar data sets, and the lack of a unified processing framework, we first used dimensionality reduction and visualization to analyze potential deficiencies caused by the incompleteness of current data set extraction methods. We found that the limited use of non-pulsar data may lead to bias in the result, which may limit the generalization ability. Based on the dimensionality reduction results, we propose a Grid Group Uniform Sampling (GGUS) method. This data preprocessing method improves the performance of Random Forest, Support Vector Machine, Convolutional Neural Network, and ResNet50 models on Lyon' s features, diagnostic plots, and period-dispersion measure (period-DM) plots in the HTRU1 data set. The average recall increased by approximately 0.5%, precision by nearly 2%, and F1 score by around 1.2% for all models and in all data sets. In the period-DM plots testing, the high-performance ResNet50 algorithm achieved over 98% F1 using random sampling. GGUS demonstrated further improvements in this test, enhancing the average F1 score, precision, and recall by approximately 0.07%, 0.1%, and 0.03%, respectively.

Key words: (stars:) pulsars: general -methods: data analysis -methods: statistical

1. Introduction

In recent years, the rise of artificial intelligence has led to the application of machine learning methods across a wide range of research and industries, aiming to solve frontier problems and drive innovation in both projects and technology (Camastra & Vinciarelli 2015; Zantalis et al. 2019). In the field of astronomy, for example, machine learning has been widely used in pulsar candidate identification (e.g., Wang et al. 2018). Given the significance of pulsar research in detecting gravitational waves (e.g., Abbott et al. 2017; Xu et al. 2023), navigating in deep space (e.g., Deng et al. 2013), and providing an exceedingly accurate timing system (e.g., Hobbs et al. 2020), discovering new pulsars is also one of major focuses (e.g., Nan et al. 2006; Wang et al. 2023). Identifying new pulsars requires filtering genuine pulsar signals from large numbers of candidates from observational data processing results, making the application of machine learning algorithms to assist in candidate identification essential.

Different from typical computer vision tasks, pulsar candidate data sets are characterized by class imbalance, with pulsars representing only a very small fraction of the data, while the majority consists of background noises and radio frequency interferences (RFI). Special algorithm designs are needed to find pulsar signals and make the algorithms more robust and transferable. Recent advancements in image recognition algorithms have led to notable successes in pulsar candidate identification. For instance, Yin et al. (2022) achieved 100% accuracy on the High Time Resolution Universe (HTRU) survey data set using Generative Adversarial Networks (GANs) and Residual Neural Networks (ResNet). Liu et al. (2024b) employed a semi-supervised method to reduce manual labeling efforts, which achieved over 90% accuracy on both the FAST and HTRU data sets with 100 labeled samples. Liu et al. (2024a) addressed the imbalance in pulsar data by using ResNet, achieving over 97.5% performance across various metrics on both FAST and HTRU data sets. These studies show the challenges of data imbalance and potential labeling issues in pulsar searches.

Consequently, many researchers have explored data augmentation techniques to expand pulsar candidate data sets to counteract data set imbalance problem. The augmentation ratios vary widely, from several times (e.g., Liu et al. 2023) to dozens of times (e.g., Agarwal et al. 2020; Liu et al. 2024a), and even up to 50 times (e.g., Wang et al. 2019). While there is no precise metric to determine when data augmentation may lead to overfitting, extreme levels of augmentation may be hazardous. Excessive data augmentation may result in models learning limited characteristics of the specific data set, resulting in a lack of generalizability (Shorten & Khoshgoftaar 2019).

Analyzing existing data sets can give us an understanding of the distribution of pulsars and non-pulsars, as well as the impact of non-pulsar signal selection on algorithmic outcomes. Moreover, when dealing with large-scale candidate data sets, it is necessary to evaluate whether the training data set is sufficient to ensure the algorithm’ s robustness and generalization capability. To help algorithms use the pulsar and non-pulsar data sets comprehensively, we proposed a method called Grid Group Uniform Sampling (GGUS) for data extraction.

Section 2 introduces the characteristics of pulsar candidate data sets. Section 3 explains the rationale for dimensionality reduction and presents a comparison of the results in the HTRU1 data set. Section 4 proposes and refines the GGUS method for training and testing data extraction. Section 5 presents the experimental results, and Section 6 discusses our results, summarizes our findings and future research directions.

2. Pulsar Dataset

In the field of pulsar candidate classification, data sets like HTRU1 (Morello et al. 2014), HTRU2 (Thornton 2013), and FAST (Wang et al. 2019) are two-class classification data sets with pulsars and non-pulsars. In this paper, we selected the HTRU1 data set, one of the most commonly used data sets, for analysis. The HTRU1 data set comprises a subset of preprocessed data from the HTRU survey, including 1,196 pulsar instances coming from 521 pulsars and their harmonic data, along with 89,996 non-pulsar instances.

Based on Lyon et al. (2016)’ s review of pulsar candidate identification and recent studies in the field (e.g., Yin et al. 2022; Liu et al. 2024b), pulsar data sets for machine learning classification can be divided into two types: numerical data based on pulsar features and visual image data sets.

The feature-based numerical data are derived from parameters that have been designed by researchers (e.g., Lee et al. 2013; Morello et al. 2014; Tan et al. 2018). Parameters such as the signal-to-noise ratio (S/N), pulse profile, dispersion measure (DM), pulsar period, and acceleration search are commonly selected as features for analysis. Among the manually designed pulsar features, we selected the features proposed by Lyon et al. (2016) as one of the test data sets in this study. Details of Lyon' s features are presented in Table 1.

Pulsar candidate samples in image format provide data details, with each feature value that can be learned by algorithms. In this article, we selected the pulsar diagnostic plots and the period-dispersion measure (period-DM) plots for analysis and testing.

The pulsar diagnostic plot consists of four sub-plots. Figure 1 illustrates diagnostic plot examples of a pulsar signal, a typical RFI, and a background noise. As one of commonly used data set formats, researchers such as Zhu et al. (2014), Wang et al. (2019), Guo et al. (2019) selected and extracted specific features from these diagnostic plots for training and testing in their algorithms.

The period-DM plots, shown in Figure 2, reflect how the S/N of a signal varies with different periods of folding and different dispersion values. In period-DM plots, the most notable distinction between RFI and pulsar signals is that RFI usually does not exhibit significant dispersion, and its S/N often peaks at zero dispersion. In contrast, noise typically shows relatively low S/N values and lacks distinctive periods and dispersion characteristics.

This paper will apply Lyon' s features, diagnostic plots, and period-DM plots obtained from HTRU1 for processing and testing.

3. Data Preprocessing and Visualization

Wang et al. (2018) argued that although manually designed features are compact and concise, their reliance on human design may lead to bias. In contrast, image-based data sets tend to yield more precise results, though they are large and challenging to train. We also found that, with increasing computational power advancing, most recent approaches are result-oriented (e.g., Lyon et al. 2016; Wang et al. 2019). However, few researchers (e.g., Wang et al. 2019) have conducted in-depth analyses of the intrinsic features of pulsar data. Hence, the first objective of this paper is to analyze the features of pulsar data sets from a macro perspective. We aim to provide a comprehensive and quantitative understanding of the overall characteristics of pulsar data.

For algorithm selection, dimensionality reduction (Garzon et al. 2022) is chosen as the basis for visualization analysis. Compared to the original data set, dimensionality reduction offers advantages such as improving data set usability, reducing computational overhead, removing noise, and making results easier to interpret and visualize (Garzon et al. 2022).

Common dimensionality reduction algorithms include Principal Component Analysis (PCA; Abdi & Williams 2010), Linear Discriminant Analysis (LDA; Tharwat et al. 2017), Local Linear Embedding (Roweis & Saul 2000), Multidimensional Scaling (Torgerson 1952), t-distributed Stochastic Neighbor Embedding (Van der Maaten & Hinton 2008), and Isomap (Tenenbaum et al. 2000), which cover various methods including linear, nonlinear, and manifold learning approaches.

When dealing with large data sets, nonlinear and manifold learning algorithms experience exponential increases in computational complexity as the data set size grows. This phenomenon may lead to the "curse of dimensionality." Therefore, efficient and interpretable linear dimensionality reduction algorithms are preferred. Among linear algorithms, we selected PCA as the method for dimensionality reduction and visualization. PCA is an unsupervised method, which can reduce data dimensions while preserving essential information, facilitating subsequent visualization and analysis.

3.1. Principal Component Analysis (PCA)

PCA is one of the most well-known linear dimensionality reduction algorithms. It projects high-dimensional data onto a lower-dimensional subspace while retaining most of the variance in the original data. The core idea of PCA is to identify the directions in which the variance of the data is maximized and project the data onto these directions, which are known as principal components (Abdi & Williams 2010).

Let the original data that needs to be reduced be represented as an $m \times n$ matrix, where m represents the dimensions of each data point, and n is the number of data points. The goal of the PCA algorithm is to reduce the m-dimensional data to k-dimensions ($m > k$).

The PCA algorithm (Abdi & Williams 2010) can be divided into the following steps:

1. Perform decentering to make the data set' s mean zero:

   $$\mathbf{X} = \mathbf{X} - \bar{\mathbf{x}}$$

   where $\bar{\mathbf{x}}$ is the mean of the data.

2. Compute the covariance matrix:

   $$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$$

3. Find the eigenvectors and their corresponding eigenvalues $ = (