Pulsar Star detection

Nirav Kamani
Department of Computer Science and Engineering
Nirma University
Ahmedabad, India
20bce116@nirmauni.ac.in

Kartavya Mandani
Department of Computer Science and Engineering
Nirma University
Ahmedabad, India
20bce120@nirmauni.ac.in

Karan Shah
Department of Computer Science and Engineering
Nirma University
Ahmedabad, India
20bce118@nirmauni.ac.in

Lakshit Kava
Department of Computer Science and Engineering
Nirma University
Ahmedabad, India
20bce124@nirmauni.ac.in

***Abstract*—Pulsars are a rare type of Neutron star that produce radio emission beam which is detectable here on Earth. They are of considerable scientific interest as probes of space-time, the interstellar medium, and states of matter, testing general relativity, gravitational wave and a lot more. As pulsars rotate, their emission beam moves across the sky, and when this crosses our line of sight, which produces a detectable pattern. As pulsars rotate rapidly, this pattern repeats periodically. Thus pulsar search involves looking for periodic radio signals with large radio telescopes. Machine learning tools are now being used to automatically label and classify pulsar candidates for rapid analysis. Classification of the candidate data sets as binary classification problems. In this paper, we have described the process of our data operations on the Kaggle Predicting-Pulsar star data-set, and various models are been trained on the transformed data-set in which SVM is most performs best with an accuracy of 97.9% on test data.**

***Index Terms*—Pulsar, Machine learning, Integrated profile, Depression measure, SNR, SVM, Classifier.**

I. Introduction

Pulsars are fast-rotating Neutron stars, which are sources of 'Quasi-regular pulses' that spread over the Radio spectrum (Giga Hertz Broad Band). Each pulse is added up to a mean integrated profile. Each pulse is almost looking similar or says the standard deviation with a mean of significantly less. The difference between pulses is measured due to various factors such as the pulsar's spin rate, rate of energy transfer to rotational motion, star-Quakes etc.

Pulsating stars are created when the core of the star is collapsed into a neutron star and the outer layer explodes into a supernova. Pulsars have a strong magnetic field and spin rapidly due to that Giga Hertz's Radiation beam being ejected from the magnetic poles of the stars. The mass of the pulsar is around 1-1.5 $M_\odot$ ($M_\odot$ is the mass of the sun), but the diameter of the star is 25km only, Which creates highly dense material like the nucleus of the atom. The density is almost the same if we compress Mt. Everest into a small backpack. The surface gravity of the pulsar is 200B times than the earth has. So it has extreme characteristics. When the rotating start immediately left the mass, the core's spin becomes much faster due to the conservation of Angular momentum. Star has some deviation between the rotation axis and magnetic dipole axis. Due to all the conditions, magnetic fields are highly dense, and a rotating magnetic field creates a beam of radiation from the poles of the star. if the earth is located in sight of view, we can observe the radio wave pulses, which create a 'lighthouse effect' while observing. Normally The frequency of spin is 0.12Hz and 642Hz. The simplified structure of pulsars is described in figure 1 taken from [1].

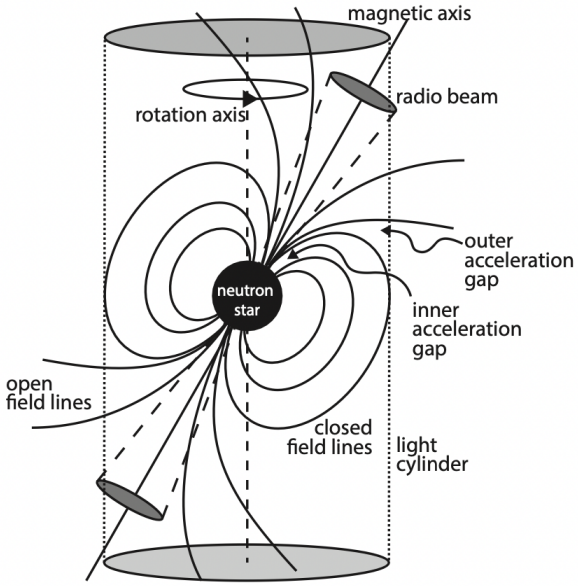

Fig. 1. Simplified schematic model of a pulsar [1]

The data set is taken from the Kaggle platform collected during High Time Resolution Universe Survey(HTRU).In this data-set, the legitimate pulsar examples are a minority positive class, and spurious examples are the majority negative class. The data set shared here contains 16,259 examples caused by RFI/noise and 1,639 real pulsar examples. These examples

have all been checked by human annotators. Each row lists the variables first, and the class label is the final entry. The class labels used are 0 (negative) and 1 (positive). The attributes are discussed in the following section.

II. Data-set Attributes

Predicting the pulsar depends on a variety of attributes or features but the most common features are the integrated profiles curve and DM-SNR curve which is compiled by various other impacting features. By taking the mean, standard deviation, Excess Kurtosis and skewness of both curves are taken into account.

A. Pulsar Profile Curves

1) integrated profile: If we see each pulse signal during each rotation of the pulsar ejected by the pulsar, it varies from pulse to pulse. So for the analysis integrated profile of the pulsars is been used Integrated profiles are seen to be consistent from different observations and generally stable in time. The process of generating an integrated profile is covered in figure 2 taken from [1]. First, the base band signal is been digitalize and converted into a time-series polarization plot. Then the de-dispersion of the signal is applied, and the frequency domain is been divided into smaller bins of frequencies. The smaller frequency spectrum is then de-dispersed to generate high power signal which is been lowered by dispersion due to ISM, now the time series pulse width is narrowed and overall power is been accumulated. After de-dispersion, Period folding is applied to the time series, trying to fill these gaps on sparsely sampled data points. Time series is now folded by the known period of the pulsar. This is been done using Modified Julian Date (MJD) to compensate with down spin and doppler shifts. Period folding will increase the power of the time series signal and eliminate the deviations, with noise at constant. Hence the Signal to Noice(SNR) is increased. The SNR has now increased by down sampling the signal, dividing the signal into smaller bins and then taking the average data of each bin similar to a fast Fourier transformation. The white Gaussian is converged to a mean value and the signal is converted to an integrated profile [1].

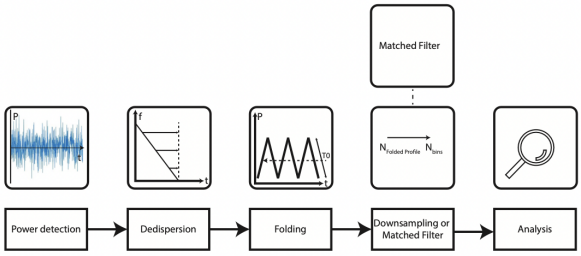

Fig. 2. Convert raw Signal Data to Time series data [1]

2) DM-SNR Curve: The Radio signals are travelled through the interstellar medium(ISM), by which the signal has fainted, so the movement causes distortion and noise will be added. As the pulsars are located far away from the earth the waves coming from them is considered plane wave with multiple frequencies, because of the ionized component of ISM the plane has frequency dependency [1]. The frequency dependency will cause a larger frequency to come earlier than with a lower frequency. The amount of pulsar signal dispersion is defined as the dispersion measure (DM). By the time the pulses spread into time and overall amplitude are decreased. How the spectral SNR is variable with a wide range of DM trails is taken into consideration by the SNR-DM curve [1].

B. A statistical measure of the curve

1) Mean and standard deviation: Mean is an essential concept in mathematics and statistics. The mean is the average or the most common value in a collection, which is the central tendency of the distribution. A standard deviation is a measure of how dispersed the data is with respect to the mean. A low standard deviation means data are condensed around the mean, and a high standard deviation indicates data are more separated.

2) Kurtosis: Kurtosis is defined as the "peakedness" of the distribution. Take some of the density from shoulder areas of the distribution, by which the curve becomes peaked at the centre and flattered at the tail. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis have heavy tails or more amount of outliers. Data sets with low kurtosis tend to have light tails or fewer outliers. This indicates how the data-set is been closely separated from each other.

3) skewness: As kurtosis is the absolute measure of outliers distribution, skewness indicates in which direction the outliers exist i.e it gives the direction of the outliers. if it is right-skewed, most of the outliers are present on the right side of the distribution while if it is left-skewed, most of the outliers will present on the left side of the distribution.

C. Machine learning Methodologies

1) Logistic regression: Logistic regression is the classification algorithm used to label and strongly binary classify input data to discrete classes. Unlike linear regression, Logistic regression uses the sigmoid classification activation function to add some of the non-linearity to the relation. But before activation, the correlation between the input and the output should be linearly variate. So the classification of the textual data model has to take many assumptions on the data-set before the training, hence the model is leaner towards the under fitting.

2) Naïve-Bayes Classification (NB): Naïve-Bayes Classification is ML supervised learning algorithm that is used to classify dependent events. Naïve-Bayes is based on the probabilistic Bayes theorem, which assumes that the events are independent of each other and all the events should be mutually exclusive. This principle can be applied to classify spam emails because the frequency of words plays an important role, and the occurrence of one word in spam email has a huge impact on another. As the algorithm

is dependent on frequency, the place of the word does not matter, though it is one of the most used spam classification techniques. NB classification calculates the probability of whether each class here is pulsar or not. And which class having the largest probability is the class of that event and provides the accurate result on classification.

3) Support Vector Machine (SVM): SVM is one of the popular classification models. As the name suggests SVM needed support vectors that create a linear Hyperplane that divides the data-set into classes. The support vector are been chosen as the maximum margin classifier. In some cases, the data is not linearly separable by linear hyperplane so Kernel functions are been used to introduce some non-linearity. The kernel function maps or transfers the whole input space to a higher dimension in which the data has been separable by Hyperplane. Linear, RBF(radial basis function), sigmoid and polynomial Kernels are been used. The SVM is then descent fast because it depends only on the vector dot product only, so in some less featured data-sets, the SVM is faster than Artificial Neural Network (ANN).

4) Decision Tree: As the NB is this classification technique also does not require prior domain knowledge. This is based on how humans can take decisions. The data-set is been split by the feature which has the current highest information of the data-set, the tree has been developed till the unit or close to a unit decision is arrived. The information gained measures how the impurity of the data-set is been decreased. Entropy is been the measure of impurity. After the tree has been created the model is ready to take decisions on new unknown data. In some cases only developing a tree is may reduce the accuracy, to improve the performance pruning on the tree i.e. cutting some of the branches are been performed during or after the tree's growth. Entropy can be found by the following equation.

$$E(D) = \sum_{I=1}^{c} -p_I \log_2(p_I)$$

$$pi : probability\ of\ I^{th}\ class$$

$$D : Data - Set$$

$$C : degree\ of\ class$$

III. Evaluation Matrices of model

To evaluate the machine learning model and how the model generalizes the data set, evaluation matrices are needed, which include accuracy, precision, recall, and F1-Score.
True Positives(TP) are examples which correctly classify the true classes, True Negatives(TN) are examples which correctly classify the false classes, and False Positives(FP) are examples which incorrectly classify the false classes i.e the actual example belongs to a false class which is classified as true, False Negatives(FN) are examples which incorrectly classify the true classes i.e. the actual example belongs to a true class which is classified as false.

	Accuracy	Recall	Precision	F1 Score
DT Train	1.0	1.0	1.0	1.0
DT Test	0.966	0.858	0.791	0.823
RF Train	0.977	0.806	0.933	0.65
RF Test	0.976	0.814	0.909	0.859
LR Train	0.976	0.787	0.947	0.860
LR Test	0.976	0.788	0.938	0.856
NB Train	0.942	0.839	0.644	0.729
NB Test	0.941	0.849	0.635	0.726
SVM Train	0.978	0.812	0.945	0.874
SVM Test	0.979	0.826	0.940	0.879

TABLE I
Result Summary Table

Accuracy is defined as,

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is defined as,

$$p = \frac{TP}{TP + FP}$$

Recall is defined as,

$$r = \frac{TP}{TP + FN}$$

F1-score is defined as,

$$f1 = \frac{2pr}{p + r}$$

1) Random Forest: A random forest is a group of different decision trees having different sizes and classifiers. The input feature are been randomly selected, which reduce the generalized error. The result of the model is been predicted by the most predicted class.

IV. Data Pre-processing

The test data has a total of 12528 number of samples in which 11275 are Negatives indicating non-pulsar objects and 1153 are positives indicates pulsars. The Data contains null values in some of the attributes which are been replaced by the corresponding median of the existing not-null values. for evaluating the models training data is divided into two sections, 80% of the training data used for fitting and 20% of it used for testing, furthermore the training is evaluated by 10 fold Cross Validation (CV) set. The specific attributes are visualized by the box plots

V. Model Training and Evaluation

For Training the data we have plot the box plots of the attributes shown in figure 3, which indicates the distribution of values. some of the attributes are closely distributed and some are sparsely. In result of that feature scaling is performed on the data-set.

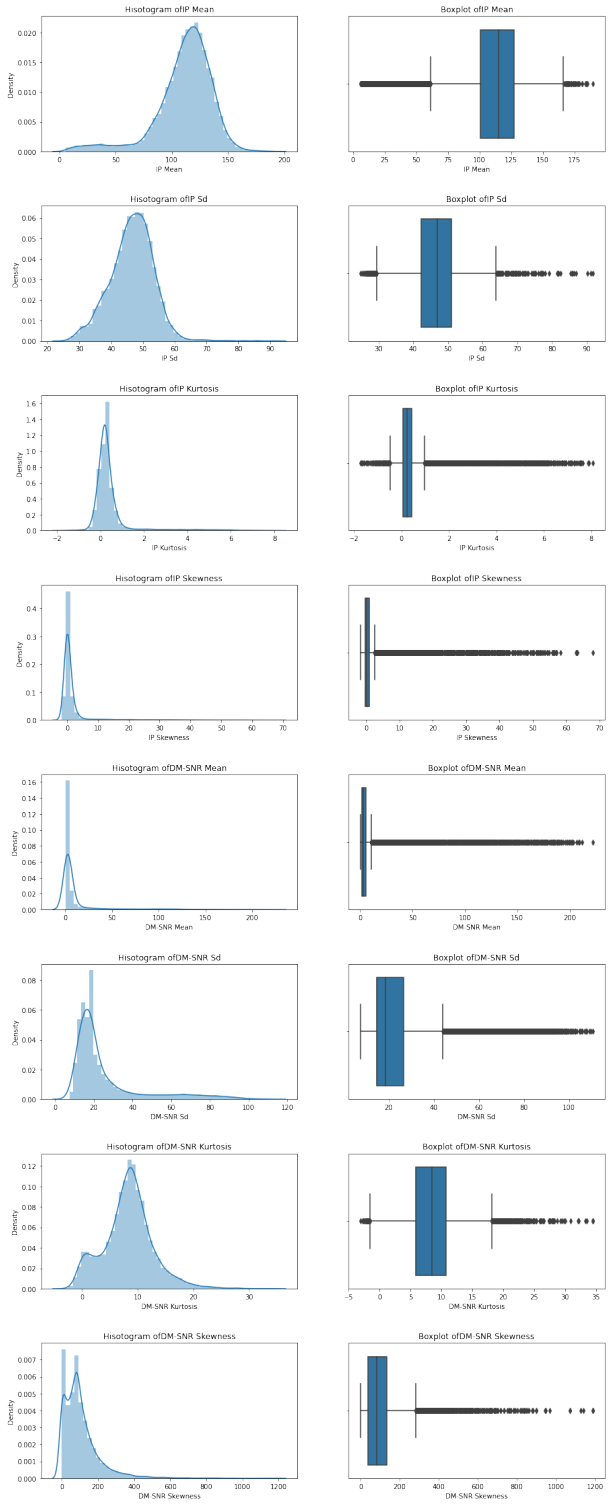

Fig. 3. Spreads of the Attibutes

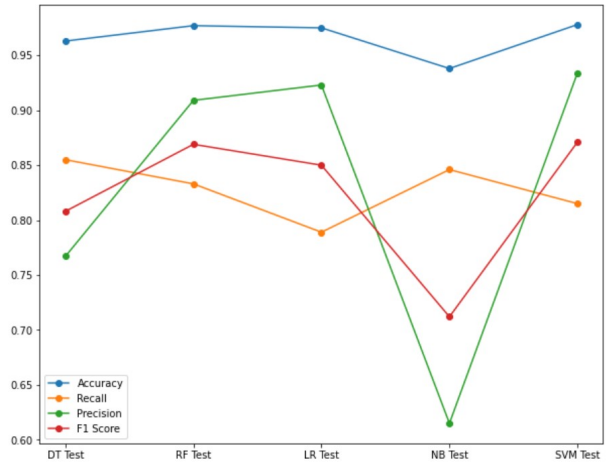

Fig. 4. Model score evaluation

We have used SVM, logistic Regression, Decision Tree, Random Forest and Naïve-Bayes model for training and validating. The train and test results are shown in table 1 and charts are shown in figure 4. Among the models SVM is fitting best on the available data-set with 97.9% on the validation testing.

VI. Conclusion and Feature Work

The model can be improve if new features are been added to the data-set but pulsars are rare astronomical objects, so data collections is might be an issue. So instead of working on Supervised learning we might hybridized the approach to semi-supervised or unsupervised learning machine learning method.

References

[1] Roelof G, Mark B and at el. Detection of Dispersed Pulsars in a Time Series by Using a Matched Filtering Approach.Master thesis Telecommunication Engineering August 2016.