March - April 2020
ISSN: 0193-4120 Page No. 2239 - 2244

# K-Nearest Neighbor for Identification of Pulsars in Astronomy

R. B. Dhumale
*E&TC Department Sinhgad College of Engineering* Pune, India rbd.scoe@gmail.com
Pushkaraj Sadegaonkar
*E&TC Department Sinhgad College of Engineering* Pune, India pushkaraj.sadegaonkar@gmail.com
N. R. Dhumale
*E&TC Department Sinhgad College of Engineering* Pune, India nrd.scoe@gmail.com

***Article Info***
***Volume 83***
***Page Number: 2239 – 2244a***
***Publication Issue:***
***March - April 2020***

***Article History***
***Article Received****: 24 July 2019*
***Revised****: 12 September 2019*
***Accepted****: 15 February 2020*
***Publication****: 18 March 2020*

**Abstract**
In radio astronomy, pulsar detection is not an easy method because most radiation detectors have telescopic noise. One of the methods of cognitive demand is to decide on the proper radiation to be released by the pulsar. In this paper, the K-Nest Neighbor based classifier has been implemented to identify pulsar stars by noise by classifying pulsar candidates from non-pulsar candidates. The nearest-neighbor concept is explained with an assortment of data. The algorithm is implemented using four statistical values of two input features. The statistics are standard deviation, excess kurtosis and skewness. The input features are integrated profile and DM-SNR curve. 17,897 observations are used to train the classification, and the average accuracy obtained is 96.54%. A detailed explanation of the algorithm is also given.

**Keywords** – *pulsars, radio astronomy, K- nearest neighbor, candidates..*

## I. INTRODUCTION

Identifying a given candidate as a pulsar is of great concern in astronomy. Since these massive neutron stars formed after the collapse of huge stars, their study has uncovered many secrets of the known universe. Through their study, astronomers are able to understand the nature of gravity, the structure of the inland media, the evolution of the universe, and so on. These pulses emit beams of electromagnetic radiation while moving. These rays are detected when the rays are directed toward the earth. The potential emission of a candidate pulsar is. Choosing the right candidate is a difficult process as there are a large number of emissions that are not pulsar and are noisy factors. Very few candidates look for Pulsar. Their identities began in the late sixties and since then the techniques for finding them have advanced greatly. Nowadays, *Artificial Intelligence (AI)* techniques are used in their detection process [1]-[4].

*Machine learning (ML)* is a subject of AI that involves learning data algorithms that learn from the data by gaining important insights and learning as a human child, of course, by experience. An important application of ML is the classification of data. Several ML algorithms for classification include *K-Nearest Neighbor (K-NN), Dissection Trace, Name Bias, Logistic Regression and Support Vector Machines (SVM).*These algorithms can be divided as *parametric and non-parametric*. For parametric algorithms, there is a predefined model structure that states the relationship between

an input and an output. Examples include logistic regression and SVMs. On the other hand, in non-parametric algorithms no such model structure is defined a priori. The model structure is solely defined based on the data. *K-NN is an example of non-parametric classification* [5].

The K-NN algorithm is based on feature similarity. It classifies unknown data points based on how its neighbors are classified. The classification of data points depends on the class of the majority of its neighbors. K is the number of neighbors. The labels of the data points near K are searched and the labels are given at the test data points. K is often chosen odd to avoid building classes. The nearest neighbors to the data point are calculated using the Euclidean distance of the data point with other data. The smallest distance data points identify the nearest neighbor. When different values of k are selected, the class label of the test point is often changed. That is, the data points for k = 3 can be classified in class 1 and the same point when k = 5 can be classified in class 2. Therefore, the value of k is chosen by cross validation for accurate classification of data points [6].

Suppose a given data set is separated into five chunks of small data sets. Now in the first iteration, the first four chunks are considered for training and the last one is considered for testing. The accuracy of these predicted values of last data set are compared with the actual values of the data set. Similarly, in the next iteration second last chunk is considered for testing and all others for training. Again accuracy is measured for it. Similarly for five chunks five iterations are run and the value of k is picked which on average has the highest accuracy or the lowest error. This method is one of the methods to choose value of k by cross validation [7].

*The higher value of k prevents the overfitting of data*. But, if the value of k is too high or equal to the number of total training data set then the test data is barely classified by taking the mean of the data points. This means that in the data set to identify pulsars by classifying candidates, every candidate would be classified as not a pulsar because the number of non-pulsar candidates is much more than pulsar candidates. And a test data point would be classified by taking the mean of the data set which will result into a non-pulsar candidate [8].

In this paper a classifier based on the *K-NN algorithm* is used to identify a given star as a pulsar. The algorithm is implemented using the median values between the two characteristics of the candidate. Two features *integrated profile* and *DM-SNR curve*.

## II. Identification of Pulsar Candidates

Pulsar finding is extremely significant in addition to be mostly complex because there are a large number of pulsars similar to the signals detected by radio telescopes. Compared to these false signals, the true signals that are actually emitted by the pulsar are very few. The traditional approach to their identification was manual by a human observer which is time consuming and cognitive demand. Therefore, several ML algorithms were previously used to detect pulsar fast and accurately [9].

Here, KNN algorithm is implemented on the data set of candidates. The approach to implement KNN is by taking average values of the data points of two classes and classifying a new data point by comparing it with these averages. The data contains two attributes of candidates that are Integrated Profile and DM-SNR curve. Four statistics of each attribute are considered which are Mean, Standard Deviation, Excess Kurtosis and Skewness. These attributes and their outputs as one for pulsar candidate and zero for no pulsar candidate are shown in table below [10].

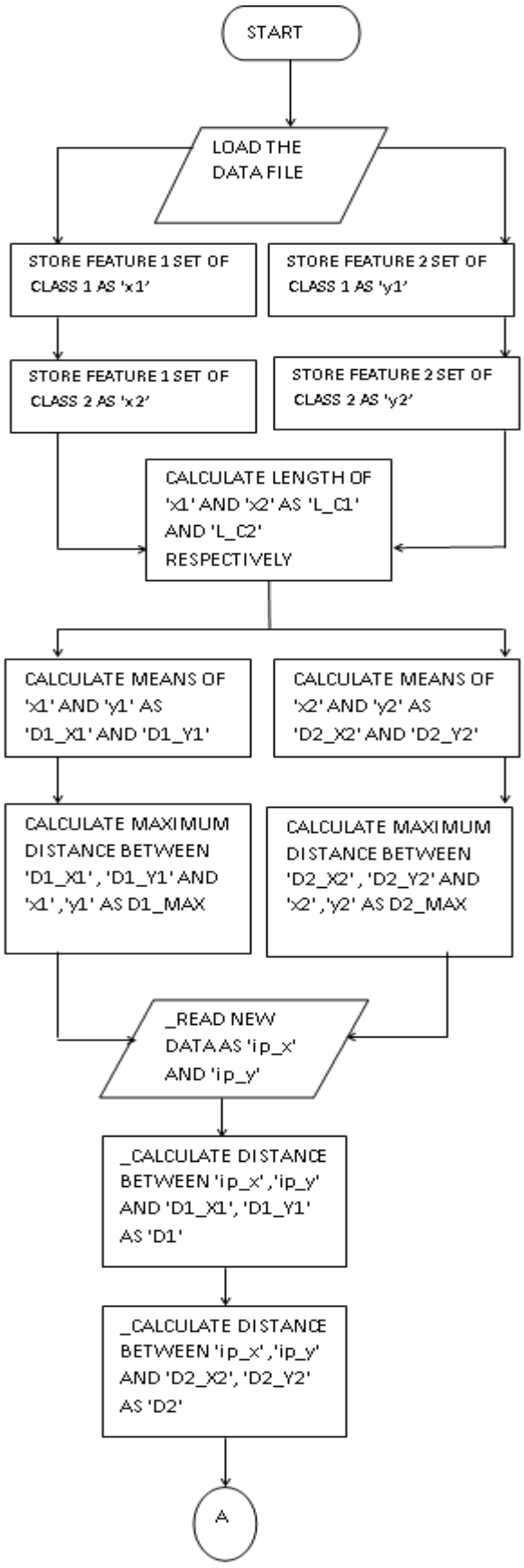
## III. Algorithm and Flowchart

The algorithm of proposed method is given below and flowchart is shown in Fig. 2

March - April 2020
ISSN: 0193-4120 Page No. 2239 - 2244

Algorithm:

1. Start.
2. Load the dataset.
3. Store features of class 1.
4. Store features of class 2.
5. Calculate mean of class 1 and class 2 data points.
6. Calculate the maximum distances of class 1 and class 2 means from the data points of respective classes.
7. Take new data to classify into group C1 or C2.
8. Calculate the distance of this new data point from the means of the two classes.
9. The smallest distance from a group is considered and the data point is classified into that group.
10. Plot the data points of both classes and also their means.
11. Plot the new data point.
12. Take the means as Centre of two circles and maximum distance of the means from the data points as radius and draw two circles.
13. Draw lines joining means of two classes with the new data point.
14. End.



March - April 2020
ISSN: 0193-4120 Page No. 2239 - 2244

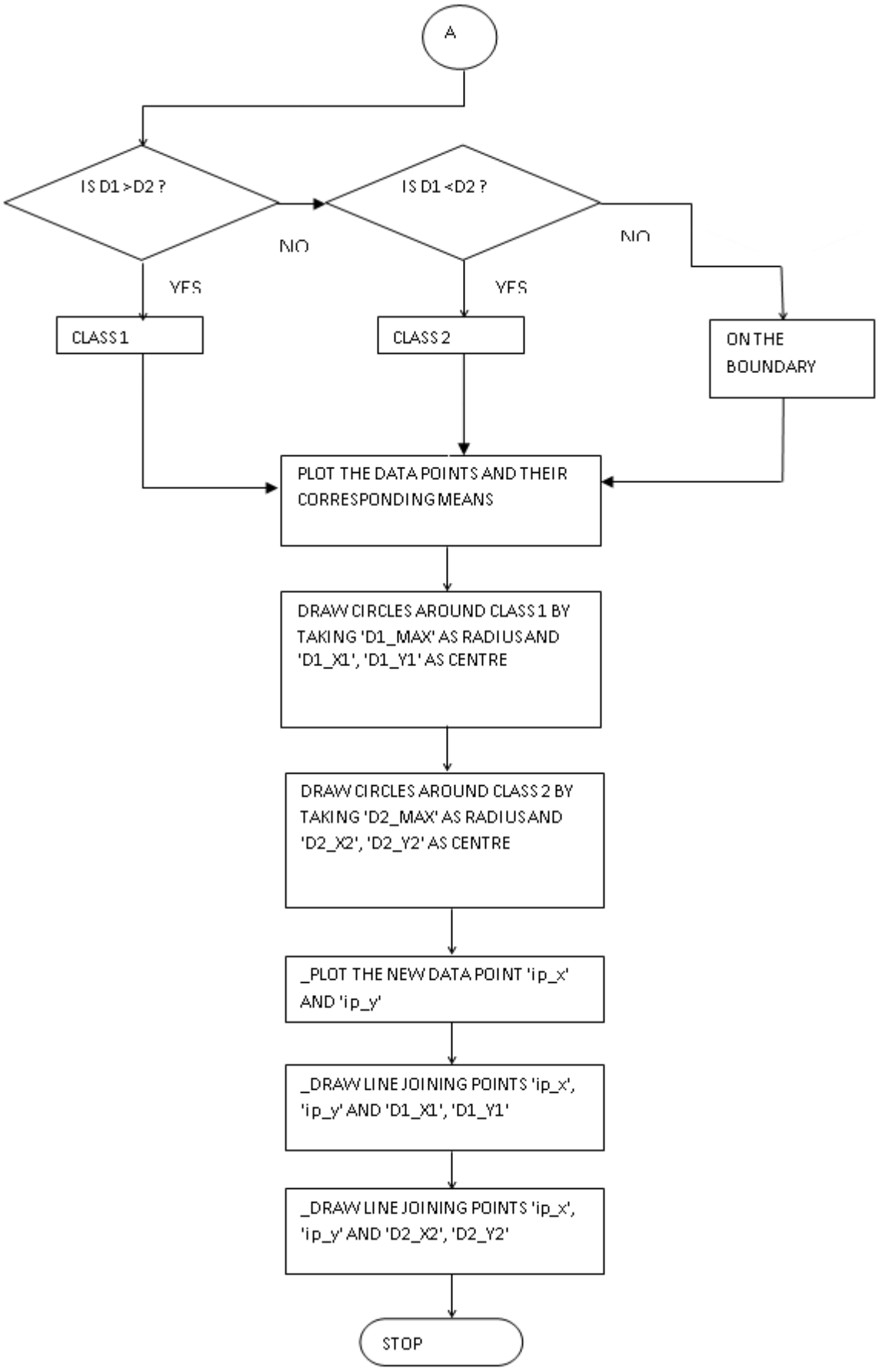



Fig. 1 *Flowchart of proposed method*

March - April 2020
ISSN: 0193-4120 Page No. 2239 - 2244

## IV. Results And Discussions

The predicting accuracy of the dataset for Linear SVM for different number of samples is calculated by cross validation with five folds. Training Time with different samples is also noted down. The Predicting Accuracy for the original dataset is 97.1% and the Training Time is 3.2886 seconds. When the number of samples taken is $\frac{2}{3}$ rd of the original, Predicting Accuracy is 96.5% and Training Time is 0.63271 seconds. Further when the number of samples taken is $\frac{1}{5}$ th of the original, the Predicting Accuracy is 96.4% and Training Time is 0.55372 seconds. The data samples are shown in Table I and Table II. Hence as the number of samples are reduced both Predicting Accuracy and Training Time decrease. Table III below shows obtained results of the experiment. The Fig. 2 shows scatter plot of the data set with only 10 samples and one statistic of the two attributes.

TABLE 1. First 10 samples of the DM-SNR Curve dataset

| Mean | Standard Deviation | Excess Kurtosis | Skewness | Pulsar |
|---|---|---|---|---|
| 3.199833 | 19.11043 | 7.975532 | 74.24222 | 0 |
| 1.677258 | 14.86015 | 10.57649 | 127.3936 | 0 |
| 3.121237 | 21.74467 | 6.896499 | 53.59366 | 0 |
| 3.642977 | 20.95928 | 6.8964 | 53.59366 | 0 |
| 1.17893 | 11.46872 | 14.26957 | 252.5673 | 0 |
| 27.55518 | 61.71902 | 2.208808 | 3.66268 | 1 |
| 1.358696 | 13.07903 | 13.31214 | 212.597 | 1 |
| 73.11288 | 62.07022 | 1.268206 | 1.08292 | 1 |
| 146.5686 | 82.39462 | -0.2749 | -1.12185 | 1 |
| 6.07107 | 29.7604 | 5.318767 | 28.69805 | 1 |

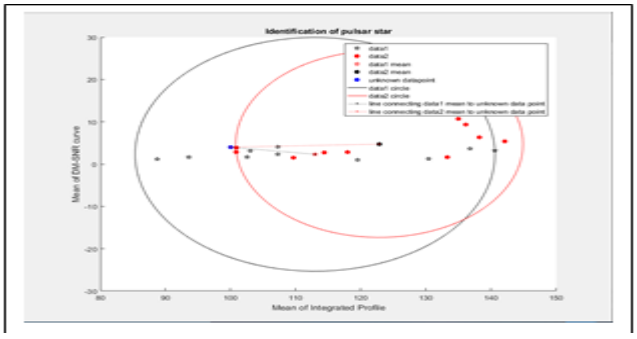



Fig. 2. *KNN with 20 samples*

The non-pulsar candidates are denoted by star and named as data1 whereas pulsar candidates are denoted by red dots and named as data2. The functioning of the KNN algorithm used in this experiment can be understood as shown in Fig.2. Black and red circles are drawn around the means of data1 and data2. Also lines are drawn connecting unknown data point with means of the data1 and data2.

TABLE 3. Predicting Accuracy and Training Time

| Samples | Predicting Accuracy | Training Time (sec) |
|---|---|---|
| 17897 | 97.1% | 3.2886 |
| 14433 | 96.5% | 0.89422 |
| 12123 | 96.5% | 0.63271 |
| 6350 | 96.2% | 0.44613 |
| 3579 | 96.4% | 0.55372 |

TABLE 2. First 10 samples of the Integrated Profile

| Mean | Standard Deviation | Excess Kurtosis | Skewness |
|---|---|---|---|
| 140.5625 | 55.68378 | -0.23457 | -0.69965 |
| 102.5078 | 58.88243 | 0.465318 | -0.51509 |
| 103.0156 | 39.34165 | 0.323328 | 1.051164 |
| 136.75 | 57.17845 | -0.06841 | -0.63624 |
| 88.72656 | 40.67223 | 0.600866 | 1.123492 |
| 99.36719 | 41.5722 | 1.547197 | 4.154106 |
| 120.5547 | 45.54991 | 0.282924 | 0.419909 |
| 27.76563 | 28.66604 | 5.770087 | 37.41901 |
| 23.625 | 29.94865 | 5.688038 | 35.98717 |

## V. Conclusions

ML algorithms have been founded as an effective approach in the field of radio astronomy for identification of celestial objects such as pulsars. In particular non-parametric supervised learning algorithm is used for classification. The KNN algorithm is studied by experimenting it with a dataset which consists of observations of pulsar candidates. The dataset was collected during the High Time Resolution Universe Survey. The

*March - April 2020*
*ISSN: 0193-4120 Page No. 2239 - 2244*

algorithm identifies a given candidate as a pulsar or no pulsar by performing binary classification on data points. The classification is done by calculating the class of the nearest neighbours. The predicting accuracy of the algorithm is estimated by 5-fold cross validation after performing fine KNN. The average accuracy of the experiment was found to be 96.54%. It is observed that as the number of observations increase the corresponding accuracy also increases. The training time is also directly proportional to the number of observations. The original dataset and the dataset with predicted values are plotted. A subset of the original dataset is plotted to explain the working of KNN.

## VI. REFERENCES

1. D. R. Lorimer and M. Kramer, 'Handbook of Pulsar Astronomy', Cambridge University Press, 2005.
2. R. J. Lyon, 'PulsarFeatureLab', 2015, .
3. R. J. Lyon et al., 'Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach', Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656.
4. M. J. Keith et al., 'The High Time Resolution Universe Pulsar Survey - I. System Configuration and Initial Discoveries',2010, Monthly Notices of the Royal Astronomical Society, vol. 409, pp. 619-627. DOI: 10.1111/j.1365-2966.2010.17325.x.
5. Ping Guo,Pulsar Candidate Identification with Artificial Intelligence Techniques., MNRAS,Vol. 1, Iss. 23, 2017.
6. R. B. Dhumale, N. D. Thombare, P. M. Bangare, “Machine Learning: A Way of dealing with Artificial Intelligence”, In Proc. IEEE Int. Conf. on Innovation in Information and Communication Technology-2019 (ICIICT 2019), St. Peter College of Engineering and Technology, Chennai.
7. R. B. Dhumale, N. D. Thombare, “Leaf Disease Classification using Machine Learning”, Proc. In International Conference of Ideas, Innovation and Impactsin Science and Technology (ICIIST-2019), Smt. Kashibai Navale College of Engineering, Pune, pp. 121, 19 March,2019.
8. R. B. Dhumale, “An Overview of Artificial Neural Networks: Part 3 Activation Functions”, CiiT International Journal of Artificial Intelligent Systems and Machine Learning, Vol. 10. Iss. 3. April 2018, pp 66-71.
9. An Introduction to Statistical Learning with applications in R by Trevor Hastie, Robert Tibshirani, Daniela Witten and Gareth James.
10. R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656.