Finding Accuracies of Various Machine Learning Algorithms by Classification of Pulsar Stars

Abhishek Seth, Arjun Monga, Urvashi Yadav, and A. S. Rao

Abstract In order to rank supervised machine learning techniques according to their accuracy, a number of them were applied on the HTRU2 dataset. Pulsars are exotic neutron stars rotating at very high RPMs which lead to a high scientific interest into recognising actual pulsars from a pool of candidates. False positives are almost indistinguishable from real positives and are generated most often due to internal and external noise and interference factors. The use of aforementioned ML techniques helps mitigate some of those problems. The raw observational data was collected by the High Time Resolution Universe Collaboration using the Parkes Observatory, funded by the Commonwealth of Australia and managed by the CSIRO. Deep investigations into the nature of exotic stars seem imminent, and a ranked list of the most accurate ML techniques presented in this paper will no doubt benefit the field of pulsar astronomy. A direct combination of future observatories and ML computation might yield unexpected results, and that is what we expect from this paper. We hope our work contributes in enabling scientific discoveries as humanity is finally becoming more and more capable of turning their heads up and understanding the mysteries of what lies beyond home.

Keywords Pulsar star · Supervised machine learning · Noise and interference · High energy radiation · HTRU2 dataset · Candidate

1 Introduction

Ever since the dawn of humanity, humans have been looking up in the sky—wondering and conjuring myths and theories to explain to ourselves and others of the happenings in the sky. This has continued till modernity, and although we have discovered far more than our ancestors ever could hope, some answers still painfully elude us. First discovered in November 1967 by Jocelyn Bell, an intermittent pulsating signal was of great interest to her and her advisor Antony Hewish. They jokingly

A. Seth (✉) · A. Monga · U. Yadav · A. S. Rao
Department of Applied Physics, Delhi Technological University, New Delhi, India
e-mail: abhishekseth.abhhi007@gmail.com

R. M. Singari et al. (eds.), *Advances in Manufacturing and Industrial Engineering*,
Lecture Notes in Mechanical Engineering,
https://doi.org/10.1007/978-981-15-8542-5_5

named the signal LGM-1 (for Little Green Men-1). Bell had discovered three more origins of pulsar in the sky, their characteristics were being investigated by the end of the year, and then, the duo published their findings in the February 1968 issue of Nature. Many more astronomers then started looking for similar signals, and thus, the science of pulsar astronomy was born.

A pulsar is a different kind of pulsar star which radiates periodic pulses of radio, X-ray, and Gamma-ray waves which are detectable on our planet. It forms after massive stars consume most of their fuel and thus their gravity overcomes the internal radiation pressure, collapsing the star. The outer layers are thus expelled to reveal a small core. Some pulsars are so because of their oscillation, but other discovered candidates also show two pulsar stars orbiting each other, i.e. a pulsar binary. During the first 15 years, after the first discoveries, the number of known pulsars grew to over 300. It is seen that the periods were distributed mainly from 100 ms to one second, and the rate of slowdown indicated that the pulsars with the shortest periods were the youngest.

A major forward step was made by the Gamma-ray observations that are re-examined using the periodicity found by radio, and many of the discrete sources turned out to be pulsating at GeV energies with pulse shapes similar to the radio profiles. Also, the X-ray pulsars were known to exist due to accretion in a binary system, of material from a star with a large and expanding envelope on to a condensed star, or an accretion disc, to temperatures around 106–107 K. Thus, the accretion process transferred angular momentum from the binary orbit to the star, spinning it up to a periodicity approaching one millisecond.

The rate of pulsar generation correlates with the supernova frequency which is one in every twenty-five to hundred years. Thus, calculating the typical lifetime of a pulsar yields the answer to be around ten million years. It loses its rotational kinetic energy through beam emission and does not rotate fast enough to emit powerful radiation. It then becomes unobservable. When a pulsar is ageing, it can still radiate in the radio domain with time periods measuring more than a second. The Crab pulsar has been around nine-hundred and sixty years, and thus, its period is comparatively short.

There is a total population of between 105 and 106 in the galaxy. Most of them are concentrated in the plane of the galaxy within a layer about 1 kpc thick and within a radial distance of about 10 kpc from the centre. The millisecond pulsars represent a smaller population of older pulsars. Their rate of "re-birth" in the spin-up process is much lower than the birth rate of the normal pulsars. They are found throughout the galaxy, but much less concentrated towards the plane than are the younger pulsars.

The Fermi LAT or Large Area Telescope as an instrument is especially useful to observe Gamma-ray emissions as it is very sensitive. The energetic light from such an emission reveals the location in its magnetosphere where the particles are accelerated, thus building a spatial model for the electric fields. It is the combination of these two techniques which KIPAC is using to describe the production of emissions within a pulsar.

Pulsar searching in modern science is done using several large radio telescopes in combination with signal processing algorithms and a touch of human ingenuity. The problem lies in the fact that rapid detection of pulsars is much more difficult because of the large amount of signals collected during a single observation. Apart from that, we have reached a level of scientific development which allows pulsars to be classified given enough time. Humans play a critical role in the identification process by recognising their emissions when they emerge.

The relevance of our and the preceding research lies in the fact that there are umpteen sources of man-made radio frequencies and noise which closely resemble the characteristic signals of a pulsar. This makes it very difficult to figure out which ones are the actual pulsars which in turn makes further research a hassle to say the least. While the earlier papers have done the bulk of the work in identifying features to be used in ML algorithms, we wondered if a higher accuracy could be achieved.

Our paper adds to common knowledge, not by trying to find more about pulsars, but by helping in classifying them from among a pool of "candidates". This is done by identifying features which are most relevant to pulsars and take the most different values for true and false positives. It is assumed that these features exist and those which fit the aforementioned conditions most closely were chosen. This classification is done using various machine learning techniques such as Naive Bayes, K-neighbour classifier, random forest, kernel SVM, and Keras classifier. Pulsars (or true positives) are assigned a "1" value, while the non-pulsars (or false positives) are assigned a "0" value. This, as is apparent, is a binary classification.

2 Problem Statement

Like most of the problems solved by technological developments, our problem is also one of reducing human labour and the need for efficiency and efficacy so that we can focus on other higher-order problems. In the recent decades, the field of radio astronomy has gained a lot of interest by amateurs and professionals alike. With the advent of highly sensitive telescopes and measurement instruments, the amount of data being captured from outer space on a daily basis has reached humongous proportions—numbers simply too huge for manual calculations.

The advent of the modern computer has relieved much of these pains, but that is just getting started. The combination of machine learning techniques with modern computers has the potential to do what has earlier been almost impossible. Due to the vast distances pulsar radiation travels, their radio, X-ray and Gamma-ray signals are severely dispersed and attenuated. It is hard to detect them but with the addition of noise and interference from natural and man-made sources, it becomes almost impossible to distinguish real pulsars from the total number of "candidates". This is our main problem, i.e. how to distinguish real pulsars from all possible candidates?

Machine learning, when implemented with properly chosen features and clean and scaled data, can be used to perform such a classification orders of magnitude faster than any human or team of humans. It becomes as simple as inputting eight numbers from observations and instantly receiving either a "yes" or "no" as the output.

It is ironic that the same technologies which helped us discover pulsars ended up distorting our observations, and another technology is helping us observe and identify them once again—this time with amazing speeds. Werner Becker of the Max Planck Institute for Extraterrestrial Physics said in 2006, "The theory of how pulsars emit their radiation is still in its infancy, even after nearly forty years of work." An optimistic view, such as the authors, might say that we did not have the tools then but do have them now.

All in all, in this paper we try to use various machine learning algorithms to make sense of the vast amount of available data on pulsar stars and effectively combat terrestrial and cosmic interference by intelligently choosing features which have the most differing values for pulsars and non-pulsar candidates. Although the identification and classification of pulsars is only one of the many problems, it is still the fundamental problem. Only when we observe them can we learn about them.

3 Research Methodology

After securing data from the HTRU2 dataset, we cleaned and scaled the data using the standard scalar from the Sklearn library. Then, we applied the following ML algorithms to the given data (obtained from the HTRU2 pulsar survey). It is prescient to have a look at the algorithms used for this endeavour. Following are brief summaries of the algorithms we have employed.

K-Nearest Neighbours (KNN)

The k-nearest neighbour algorithm is a nonparametric supervised machine learning method. This method makes no underlying assumptions about the data and thus has vast applications in real-life scenarios. While training, the algorithm learns the positions of the training points and their classifications. Then, during the actual testing, the algorithm simply calculates the relative distance between the point and the pre-established groups, and thus classifies the point to one of the groups. We used a KNN classifier to classify the candidates into pulsars and non-pulsars [1] (Fig. 1).

3.1 Random Forest

This method improved upon decision trees such that it avoids the problems associated with overfitting. Random forests are a way of averaging multiple deep decision trees. Random forests are examples of ensemble learning algorithms where multiple subsets of the original dataset are created, and many different classifiers may be applied to

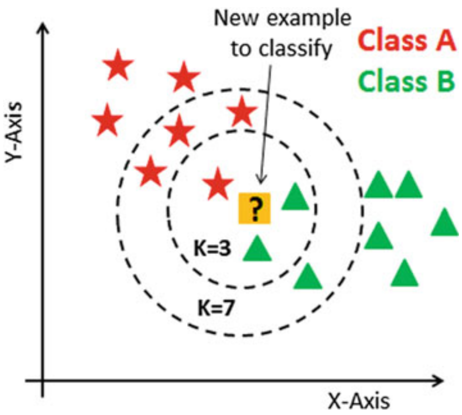

Fig. 1 Graphical representation of KNN classification [2]

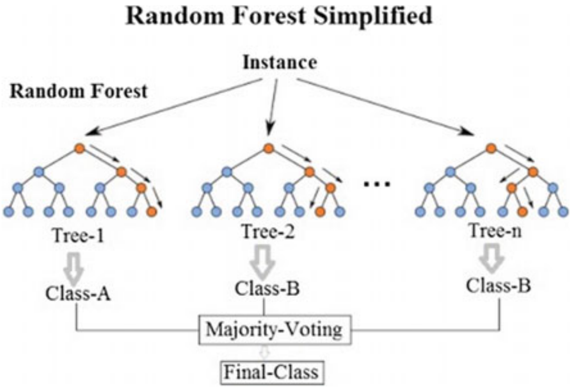

Fig. 2 Random forest algorithm [4]

each of them. In case of random forest, multiple decision trees are used and the final output is decided after taking a vote from all trees. The aggregate prediction wins [3] (Fig. 2).

3.2 Support Vector Machine

SVM (or support vector machine) method divides the data into classes using hyperplanes. First all points are plotted onto an n-dimensional hyperspace (where n is the number of features in the dataset). Then, the hyperplane which best separates the classes such that both accuracy and robustness are maximised. To prevent misclassifications, a margin is set between the hyperplane and the closest points to it. This algorithm is also robust to outliers. Linearly separable data can be directly classified using SVM, but the nonlinear data needs extra defining features to find an appropriate

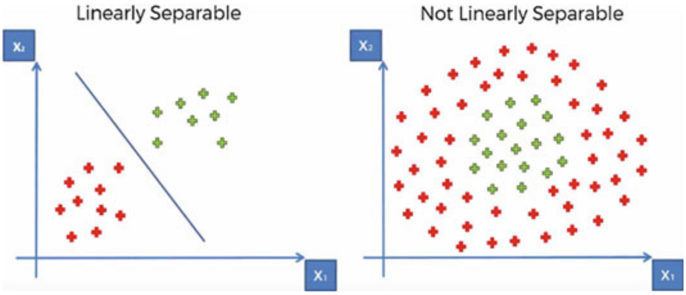

Fig. 3 Linearly and nonlinearly separable data [6]

hyperplane. The “kernel trick” is used in many types of classifiers when the data is nonlinearly separable [5] (Fig. 3).

3.3 Naive Bayes

The Naive Bayes algorithms are a family of algorithms which have one thing in common that they are based on the Bayes’ theorem. The Bayes theorem states that the a priori probability of an event given the a posteriori probability of another event is given is as follows (Fig. 4):

$$P(A/B) = \frac{(P(B/A).P(A)}{P(B)}$$

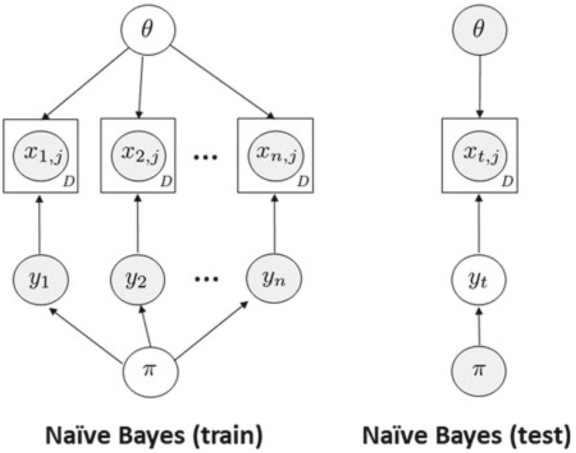

Fig. 4 Naive Bayes algorithm—testing (*R*) and training (*L*) flowcharts [7]

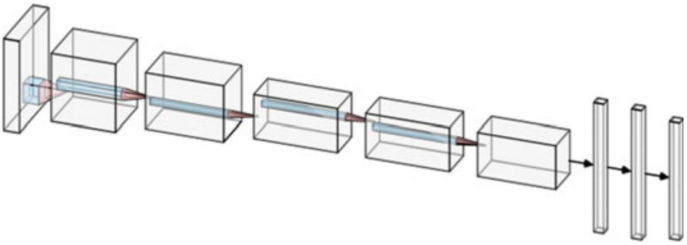

Fig. 5 Layers of a neural network [9]

The Naive Bayes assumptions are that each feature makes an *independent* an *equal* contribution to the outcome. These conditions might impose restrictions when implemented in real-life scenarios. When the independence condition holds, the Naive Bayes algorithm can be more efficient than others. It is highly scalable and gives probabilistic predictions [8].

3.4 Keras Classifier (Based on Neural Networks)

Neural networks are layers of neurons connected through synapses inside the brain. To emulate the processing methods of the brain, artificial neural networks which are analogous to biological ones were developed. Each neuron receives a signal from its predecessor, and if the signal is more than its activation threshold, the signal is propagated. Learning happens during backpropagation where the weights and biases of the individual neurons and layers are tweaked such that the next iteration will give a more accurate result. Keras is a high-level neural networks API, written in Python. Its advantage is mainly in its ability to enable fast experimentation (Fig. 5).

4 Results

After running the code for each kind of classifier, we tabulated the results. The given Table 1 shows the loss and accuracy of the model for epochs one through ten for the Keras classifier.

Plotting the accuracy and loss against number of epochs, we get the following curves (Figs. 6 and 7).

As shown, the best accuracy achieved is **97.90%** during the sixth epoch. The final accuracy achieved is **97.86%**.

Following are the final accuracies achieved using the aforementioned ML algorithms (Table 2).

Table 1 Loss and accuracy per epoch for the Keras classifier

Epoch No.	Loss	Accuracy
1	0.1001	0.9764
2	0.0905	0.9769
3	0.0847	0.9778
4	0.0808	0.9780
5	0.0789	0.9782
6	0.0770	0.9790
7	0.0762	0.9787
8	0.0752	0.9788
9	0.0744	0.9789
10	0.0740	0.9786

Fig. 6 Accuracy versus epochs

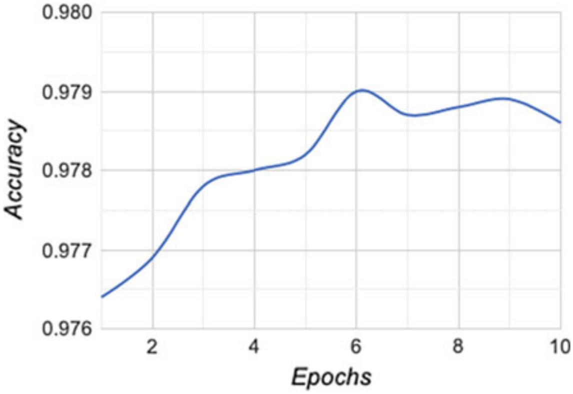

Fig. 7 Loss versus epochs

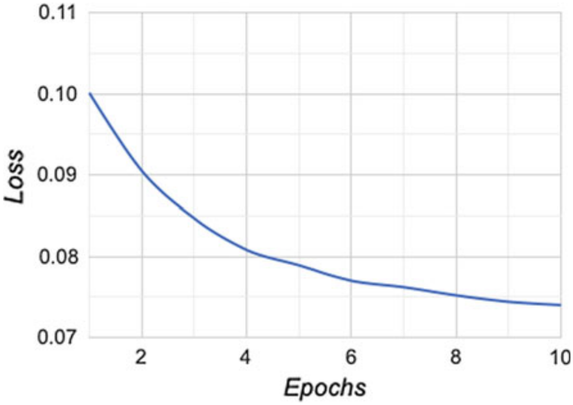

Table 2 Algorithm versus final accuracy

Algorithm implemented	Accuracy achieved (%)
K-nearest neighbours	97.944
Keras	97.900
Naive Bayes	95.061
Random forest	98.234
Support vector machine	98.122

5 Conclusion

All algorithms applied by us on the given data have given satisfactory results, which is mostly due to the high quality of the data. We found through the Keras implementation that neural networks work better for this particular set of data and parameters. As more algorithms are developed, we could hope to see higher accuracies which might ultimately do away with the need of manual classification. More study is required to understand the true nature of pulsar emissions and relates properties, but at least classification can now be done on a shorter timescale which will definitely lead to faster research.

References

1. GeeksForGeeks article. https://www.geeksforgeeks.org/k-nearest-neighbours by Anannya Uberoi
2. DataCamp article. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn
3. GeeksForGeeks article. https://www.geeksforgeeks.org/ensemble-classifier-data-mining by Avik Dutta
4. Medium article. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
5. AnalyticsVidhya article. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code by Sunil Ray
6. Medium article. https://medium.com/@ankitnitjsr13/math-behind-svm-kernel-trick-5a82aa04ab04
7. TowardsDataScience article. https://towardsdatascience.com/information-planning-and-naive-bayes-380ee1feedc7
8. GeeksForGeeks article. https://www.geeksforgeeks.org/naive-bayes-classifiers
9. StackExchange article. https://datascience.stackexchange.com/questions/14899/how-to-draw-deep-learning-network-architecture-diagrams (Answered by Pablo Rivas)