ASIPS: AUTOMATED SYSTEM FOR IDENTIFYING PULSAR STARS FROM PULSAR CANDIDATES

Rathnappuli Gamage Venoli Shehara

A dissertation submitted in partial fulfilment of the requirement for the BEng (Hons) in Software Engineering

Department of Computing

Informatics Institute of Technology, Sri Lanka

in collaboration with

University of Westminster, UK

2021

Abstract

Pulsar stars are rare type of remnants from the core of extinguished stars that emit two beams of energy per rotation. Energy beams of pulsar stars are detectable through radio telescopes. However, such telescopes also detect other signals that could be considered as noise such as background internet radio waves. Filtering pulsar signals among noise has previously been done manually. The development of telescopes resulted in an increase in the consolidation of observed data. Machine learning algorithms were used successfully to classify pulsar star candidates. Usual batch learning algorithms were effective for early surveys. Recent pulsar data collections roughly have a velocity of nearly 0.5-1 TB of information per second. Thus, the need of developing algorithms that can work with data streams was observed eliminating the need of store and classify data. Subsequently, stream classification algorithms were applied to address the issue of big data. This leads to the crux of this research.

Stream classification algorithms are specially developed to match with the characteristics of data streams such as concept drift and limited memory. Extremely Fast Decision Tree is one of the stream classification algorithms that can learn incrementally when it meets new data. Learning incrementally from a pulsar candidate stream should not be harmed by imbalances of candidate data. This dissertation is a result of the research of improving Extremely Fast Decision Tree, for imbalanced data streams. The resulted algorithm, *Gaussian Hellinger Extremely Fast Decision Tree (GH-EFDT)* is accurate, fast and avoids the pitfalls of class imbalance and concept drift. Benchmarking with existing stream classification algorithms for candidate selection, *GH-EFDT* perform better with the two standard datasets HTRU1 and HTRU2. Not only for the pulsar candidate selection, *GH-EFDT* performs equally with similar algorithms for other imbalanced data streams. The developed system using *GH-EFDT* for the candidate selection, *Asips*, is a web application with number of other features and a user-friendly GUI. Also, these features are available as an API called *Asips-for-Pulsar-Astronomy*.

Keywords: Online Learning Algorithms, Hellinger Distance, Gaussian Distribution, Extremely Fast Decision Tree, Astroinformatics, Pulsar Candidate Classification.