IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 30-35
www.iosrjournals.org

Star-Galaxy Classification of Photometric Data – A Comparative Study of Machine Learning Algorithmic Models

Shashank Shetye Saudagar[1], Sujit Gawas[2], Dattaprasad Ekavade[3], Kavita Asnani[4]
[1]*(Computer Engineering, Goa College of Engineering, India)*
[2]*(Computer Engineering, Goa College of Engineering, India)*
[3]*(Computer Engineering, Goa College of Engineering, India)*
[4]*(Computer Engineering, Goa College of Engineering, India)*

Abstract: *The classification of point source objects into stars and galaxies largely depends on spectrographic surveys, which are expensive and time-consuming. This paper attempts to estimate, with a considerable accuracy, the possibility to classify these objects using only photometric data. Various machine learning algorithms are used to determine correlation, if any, between the properties of an object and photometric parameters. The focus is on the pre-processing of data to generate the most accurate predicting feature. In-depth analysis is performed for the respective model to determine the reason to give the most accurate fit. Efficient implementation of various machine learning algorithms, representing different mathematical models is done and the performance of the same is measured and compared against each other, based on accuracy and confusion matrix.*

Keywords: *Machine Learning, Star-Galaxy Classification, Process Model*

I. Introduction

Astronomical surveys allow astronomers to look up at celestial objects and perform analysis, without spending their resources on lengthy observations. Most surveys are photometric in nature i.e. the flux or intensity of an object's electromagnetic radiation is measured in a particular band. Multi-wavelength surveys can also be done using multiple detectors, each sensitive to a different band of the electromagnetic spectrum [2, 6]. Classification of these objects is essential to carry out in-depth research. [3]This is usually done with the help of spectrographic analysis. The nature of an object is determined by measuring the bumps and wiggles in the flux of the object and correlating the same to pre-existing spectrographic fingerprints. Spectrometric Analysis, despite being an accurate method of classification, is highlytime-consuming and resource-intensive. Hence, performing spectrometric analysis of objects in large surveys, for e.g Sloan Digital Sky Survey (SDSS) is a difficult task. Even though SDSS catalogues 900+ million objects, spectrographic data of only around 3 million objects is available. Despite being an impractical option, Spectrometry is the only feasible option due to a lack of a visible structure [1].

The classification of celestial objects with a certain accuracy using low-resolution photometry is an interesting problem. Compared to Spectrometry, it is less complex and more robust, requiring minimal specialized equipment. [6] It can be done by computing correlations of the photometric data with previously classified data using the modern statistical methods viz the machine learning approach. In the approach used in this paper, analytical models are used. These models produce reliable and repeatable decisions and aid in the determination of hidden insights by learning from trends in the historical data. These algorithms are designed for operations on large datasets with multiple parameters.

However, these algorithms are based on certain assumptions. Reliable results cannot be computed using merely the raw data. Data used to construct predictors need to be complete, free of inconsistencies and aberrant data. In order to meet these parameters, the data needs to be acquired, prepared, and processed to allow the algorithms to compute as intended with maximum accuracy. The intent of this paper is to prove that photometric data is sufficient for the classification of astronomical point objects. The focus of this paper can be broken down into two points. The first point is to describe the process of data acquisition and the pre-processing of data to make it suitable for predictor construction using machine learning algorithms. The second point is the analysis and comparison of the different machine learning algorithms available for classification. These algorithms have been implemented using efficient data structures, with different mathematical models and techniques to classify the objects. These models are compared to the underlying phenomenon to determine if new insights can be discovered.

Section 2 covers the related work regarding classification of astronomical objects. Section 3 describes the proposed solution. Section 4deals with acquisition and pre-processing of data and presents the evaluation details and related observations. Section 5 covers the results and conclusion.

II. Related work

[1] describes Data Release 11 (DR11) including all data acquired through 2013 July, and Data Release 12 (DR12) adding data acquired through 2014 July (including all data included in previous data releases), marking the end of SDSS-III observing.

[3] tries to solve the problem of classification of stars into main sequence, quasars, and white dwarfs using photometric data across five different bands (u (ultraviolet), g (green), r (red), i (infrared) and z (very-near-infrared)) from DR-7 of Sloan Digital Sky Survey (SDSS). The test data consisted of 2500 quasars (QSOs), 4000 main sequence and red giant stars (m/r stars) and 981 white dwarfs (WDs). The training data consisted of 3000 QSOs, 7000 m/r stars and 3000 WDs. The feature vector was formed by taking thedifference in magnitudes between adjacent photometric bands. Visualisation of test data using bivariate color-color diagrams showed a clear distinction enabling to construct decision planes. Unsupervised Machine Learning algorithm (K-means clustering) and Supervised Machine Learning algorithms (MLR, GDA, KNN and SVM) were tested for accuracy on training and test data. K-nearest neighbours (k=4) and multinomial SVM using a Gaussian kernel (C=100) were the algorithms with the best performance.

[4] used various machine learning algorithms for multi-wavelength data classification into AGNs, stars and normal galaxies using data from optical, X-ray and infrared bands. Different algorithms like Learning Vector Quantization (LVQ), Support Vector Machines (SVM) and Single Layer Perceptron (SLP) were used for the classification. The histogram was used as the feature selection technique. In paper [5] a research work is presented on nature-inspired classification for mining social space information. As per the results, the performance of SVM models was comparable or superior to that of the NN-based models in the high dimensional space. LVQ and SLP showed better performance when fewer features were chosen.

III. Proposed method

In this paper, the evaluation of steps to be taken for the acquisition and pre-processing of data is presented. This includes data acquisition, construction of feature vector, data visualization to determine the general trends, and evaluation of outliers and other aberrant data. A comparative study of different mathematical models is done, based on the accuracy of algorithms to classify the pre-processed data. Each method is analysed to gain insights how a model can fit the data.

IV. Evaluation

The evaluation is divided into two subsections. In section 4.1,the data acquisition and pre-processing methodology is described. It includes determination of correlation and general trends, generation of feature vector and cleaning of data. In Section 4.2, a comparison of various models using different machine learning algorithms is done to find the model which generates the best predictor to classify the data into star or galaxy.

4.1 Acquisition and pre-processing of data

Sloan Digital Sky Survey (SDSS) is a multiple filter imaging and spectroscopic redshiftsurvey using a 2.5 M wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. It is one of the most ambitious projects of its time and one of the first attempts to capture the data in the sky digitally. When the data collection began in 2000, it collected more data in first few weeks than in the history of the field of astronomy. Currently, it has already accumulated 140 petabytes of data and is continuing to read and analyse data at around 200 GB per night. It contains photometric data of about 500 million objects and spectrographic data of around 3 million objects.

The data collected by SDSS is stored in a relational database on the SDSS SkyServer. The schema contains 109 tables and 59 Views defined on them. The analysis in this paper focuses and uses the data from the PhotoPrimary view, derived from PhotoObjAll table. The view contains 509 columns. This includes but not limited to quantities like components of objects velocity, Sky Flux Inverse Variance, Point Spread Function magnitude and flux, Exponential fit magnitude and flux, De Vaucouleurs magnitude and flux, Petrosian magnitude in each of the 5 bands namely U, G, R, I and Z(Ultraviolet, Green/visible, Red, Infrared and near Infrared).

The data collected by the survey can be accessed over the internet using the SkyServer. The Sky server is a web interface that provides many tools to read the data From Microsoft SQL server where this data is stored. The data can be extracted in various formats (HTML, CSV, XML, JSON, FITS). For this project, data was extracted in the CSV format. SkyServer can generate queries with amaximum output of 5,00,000 Rows and timeout of 600 seconds. Since the requirement of data was way more than that, Catalogue Archive Server (CAS)

Jobs were used. It is an online workbench for large scientific catalogues which is designed to emulate local access in a web environment. The acquired dataset had a total of 585,742,000 objects.

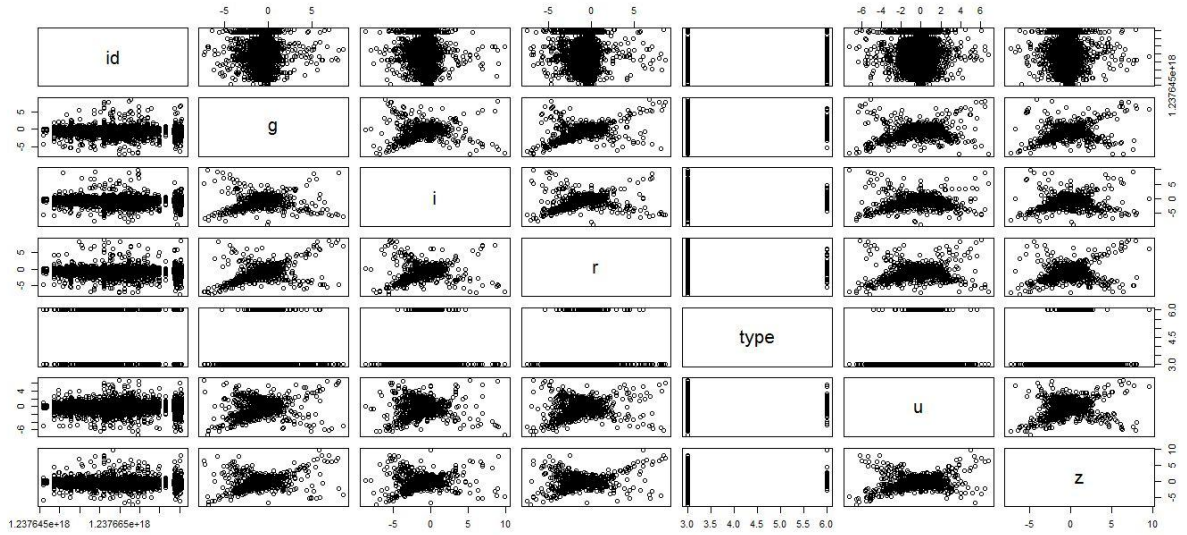


Fig. 1: visualisations of different feature vector parameters

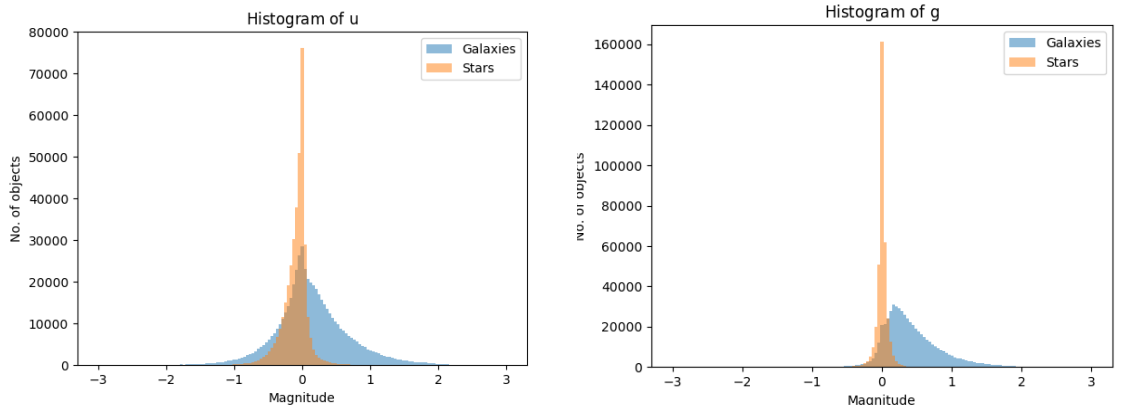
To construct the feature vector,[3] was taken as the reference. The colour indices are same as used in [3]. The approach suggested in [3] for star classification was tested and rejected as the results had an accuracy of only 61%. Post this, several parameters were tried using $model$, $cModel$, $psfModel$ and $Petrosian$ magnitudes.After the analysis of the histograms of each parameter, $cModelMag$ and $psfMag$ were chosen to build the feature. $cModelMag$ is by itself a complex parameter which combines the parameters of $de\ Vaucouleurs$ and $Exponential\ fit$ magnitudes with (1):

$$F_{composite} = fracDeV\ F_{deV} + (1 - fracDeV)\ F_{exp} \quad (1)$$

where F_{deV} and F_{exp} are the de Vaucouleurs and exponential fluxes, respectively. The coefficient (clipped between zero and one) of the $de\ Vaucouleurs$ term is stored in the quantity$fracDeV$. The PSF magnitude stands for the point spread function. The optimal measure of the total flux is determined by fitting a PSF model to the object.

The dimensionality of the feature vector is reduced using the difference of $cModelMag$ and $psfMag$as the .feature vector suggested by [2]. This quantity is also used by SDSS to determine whether an object is Star or Galaxy. [1] The data gave an initial accuracy of 82%.

The data was then visualised in a histogram for each band together and separately for stars and galaxies.It was noticed that most of the data follow a normal distribution. It was also visible that there is only a small overlap amongst the values representing stars and galaxies. These histogram plots were used to determine the threshold to determine the outliers. Fig.2 describes the histograms and thresholds for the data in each band.



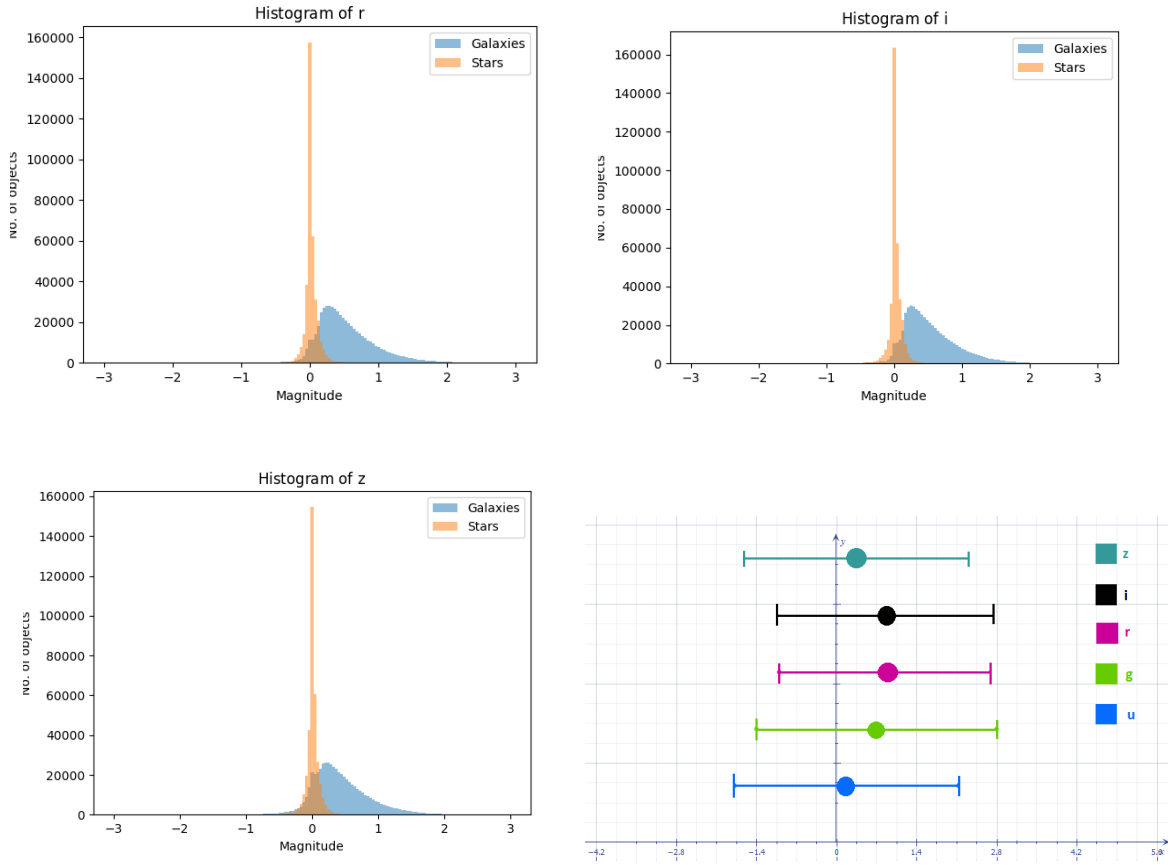


Fig. 2 – histogram and bounds of pre-processed data

The missing values in the data are denoted by -9999 value. Elements with such missing values are eliminated using the *grep* Linux command. Imprecise data and data having multiple outcomes for the same set of inputs is eliminated. Outliers were filtered using the values mentioned above. The thresholds were checked separately for all 5 different bands. Total of 8% of data was filtered from the original set.

1.2 Evaluation of Algorithmic Models

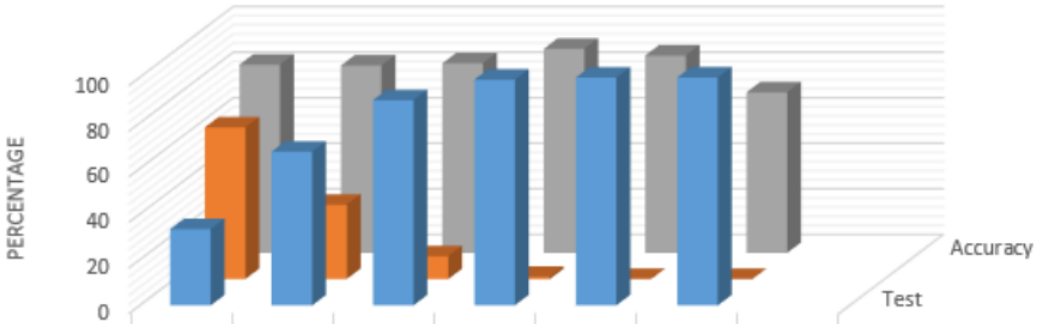
Analysis is performed on the following algorithms:

1.2.1 K – Nearest Neighbours (KNN) – This algorithmis anon-parametric method used for classifying objects, based on closest trainingexamples in the feature space. KNN is a type of instance-based learning or lazy learning, where the function is only approximated locally and allcomputation is deferred until classification. The KNN algorithm is amongthe simplest of all machine learning algorithms. In this algorithm, all objects are plotted in an N-dimensional space (where N is the number of parameters in the feature). An object is classified by a majority vote of its neighbours, with the object being assigned to theclass common among its k nearest neighbours (k is a positive integer,typically small). K = 4 was chosen, giving an average accuracy of 87.52% with raw data and 88.75% after pre-processing of data.

1.2.2 Multinomial Logistic Regression (MLR) – Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).The Principle behind the MLR is that every parameter in the feature is given a certain weight. The sum of the product of the weights and parameters is then given to the sigmoid function which determines the class of object it belongs to. For multinomial regression,the one verses all strategy was used and the prediction was chosen with minimum cost to determine the class of the object. An average accuracy of 85.15% was obtained for raw data and 87.57% was obtained after pre-processing the data.

1.2.3 Naïve Bayes Classifier - It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to

outperform even highly sophisticated classification methods. It was observed that Naïve Bayes gives a very accurate prediction even when the size of training dataset is small. Figure 3 describes the accuracy in percentages.



	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Test	33.47	67.375	89.955	99.03	99.89	99.97
Training	66.525	32.62	10.04	0.965	0.105	0.025
Accuracy	82.53	82.19	83.17	89.47	86.535	70.391

Fig. 3– evaluation of naïve bayes with different splits of training and testing sets

4.2.4 Decision Trees - A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The resulting classification tree built on the predicting model can be used for decision making. The tree is learnt by splitting the feature into subsets based on attribute value test. This process is repeated for each derived set in a recursive manner called recursive partitioning. An average accuracy of 91.06% was obtained for raw data and 92.35% was obtained after pre-processing the data.

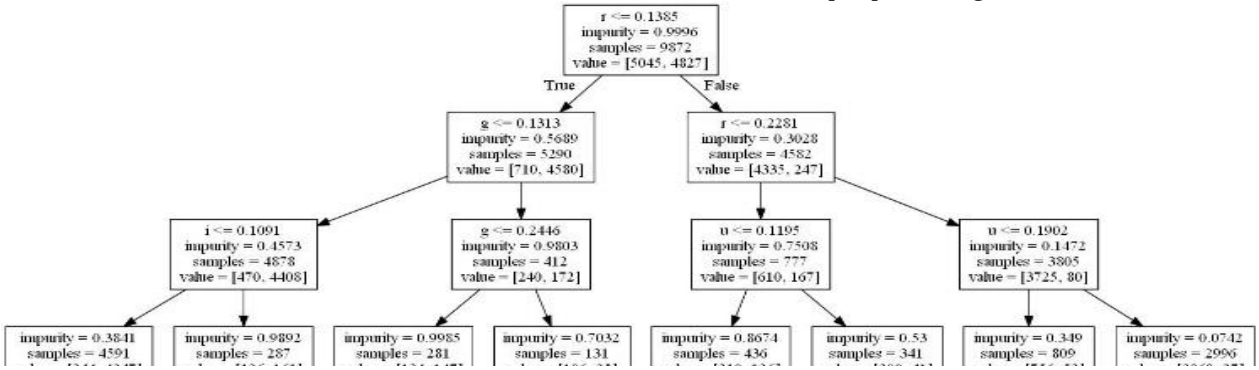


Fig. 4 – decision tree predictor

V. Results

From the analysis of the algorithms, the most accurate predictor for star-galaxy catalogue classification is constructed using a Decision Tree predictor. Use of pre-processed data leads to an improvement in each predictive model. Naïve Bayes is suitable when the training data is less as the accuracy decreases very slowly compared to the decrease in the data.

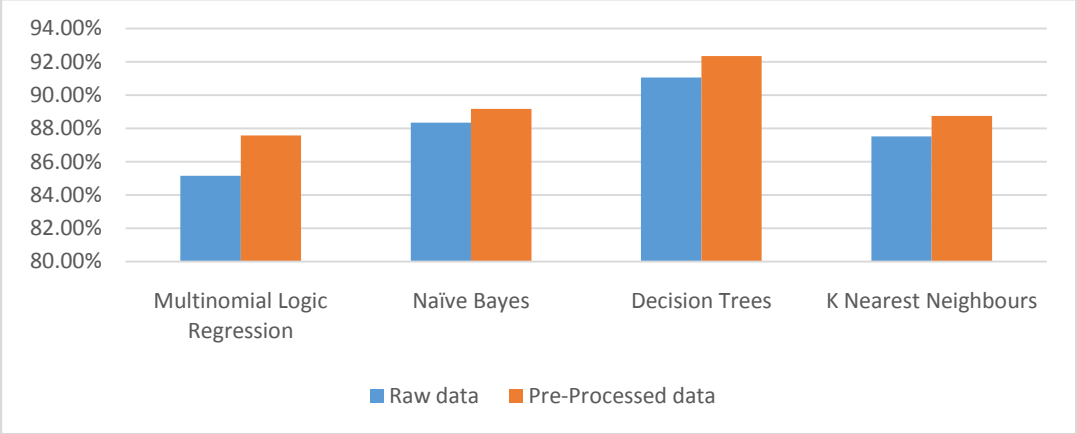


Fig. 5 – comparison of accuracy of machine learning algorithms

VI. Future Work

Randomness on the training set is an important criterion, a random forest based algorithm can be implemented, assigning a weight to predictors with different models. It has been observed that the loading time is much higher than the time taken for processing. To tackle this, an approach using distributed system can be used to access the data in parallel. In-depth research can be carried out on the entire dataset to determine the existence of more correlations between the parameters. These correlations can be added to the feature vector, thereby increasing its accuracy.

References

[1]. Alam, Shadab and Albareti, Franco D and Prieto, Carlos Allende and Anders, Friedrich and Anderson, Scott F and Anderton, Timothy and Andrews, Brett H and Armengaud, Eric and Aubourg, Eric and Bailey, Stephen and others. The eleventh and twelfth data releases of the Sloan Digital Sky Survey: Final data from SDSS-III. *The Astrophysical Journal Supplement Series219(1), May 22, 2015,1--12.*

[2]. W. Romanishin An Introduction to Astronomical Photometry Using CCDs. *University of Oklahoma (31), October 22, 2002.*

[3]. Waisberg, I. R. Astronomical point source classification through machine learning. *Stanford project year 2013., December 13, 2013.*

[4]. Yanxia Zhang, Yongheng Zhao, Automated Clustering Algorithms for Classification of Astronomical Objects, *Astronomyand Astrophysics, 422(3) · March 2004, 1113--1121*.

[5]. K Dhanasekaran and B Surendiran, Nature-inspired classification for mining social space information: National security intelligence and big data perspective, Green Engineering and Technologies (IC-GET), 2016 Online International Conference, pp. 1-6, 2016.

[6]. P. Huijse and P. A. Estevez and P. Protopapas and J. C. Principe and P. Zegers, Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases, *IEEE Computational Intelligence Magazine 9(3), 2014,27-39.*