Classification Of Galaxies

Smile Garg
Reg: 11705726
Sec: K17KH
Smilegarg1405@gmail.com

Abstract: In this paper we are trying to solve a problem of galaxy classification. In this structure of galaxy will be classified. We will apply machine learning techniques to solve this. In this project we have used both supervised and unsupervised techniques to examine the Galaxy Zoo dataset of 6,679,44 pre-classified galaxies. On this dataset we have applied multi-class classifiers using following algorithms: SVM with "Linear" kernel, decision tree, random forest etc. The classification tells whether the galaxy is spiral or elliptical or of some other shape. We have applied different algorithms to same dataset to know which algorithm gives the best accuracy. We have also used regression to check the associativity of galaxy's structure with the available classes. For visualizing the data unsupervised learning algorithms are used. These algorithms are: K-means clustering and agglomerative clustering.

I. INTRODUCTION

There are uncountable number of galaxies dispersed across the celestial spaces which describes the arrangement of the universe on the largest scales. By studying the dissemination of physical properties of these celestial bodies we can get the knowledge of past, present and future of this universe as it holds many clues from which we can get vast information about the it. Observation studies are done on these properties by cosmologists and astronomers to get knowledge and important data from these bodies and contribute some vulnerable thing into science.[1].

Within last few years, surveys of sky have done on a large scale with the advancement in technology. There are many bodies present in the space. An organisation named Sloan Digital Sky Survey also knows as SDSS registered approximately 1.2 billion bodies. The area, they have covered to find these objects was almost one third of the sky.[2]. With this number, it has shown the evidences of vastness of universe and also shows its complexity. By analysing these things, we need some analysis technique to handle and process this vast amount of data. It would be much better if the techniques will be robust and automated. As the universe is rapidly changing, these techniques should also be changed and upgraded accordingly and continuously as the new surveys (such as Large Synopsis Survey Telescope (also known as LSST)) of new generation keeps on adding new and large amount of data to the existing data.[3].

By looking at the structure of the galaxies, the astronomers try to retrieve some knowledge out of it. For example, spiral structured galaxies have cold gas feeding young stellar nurseries whereas on the other hand the elliptical shaped galaxies have old and dying stars. The distribution of galaxies according to their structure gives the information about the evolution of the cosmos, that is why mostly astronomers classify galaxies depending on their structure. For example, there is a classification technique named "Hubble tuning-fork" is also present. Its figure is shown in FIG 1. This scheme classifies galaxies whether it has spiral or elliptical or irregular(other) structure.

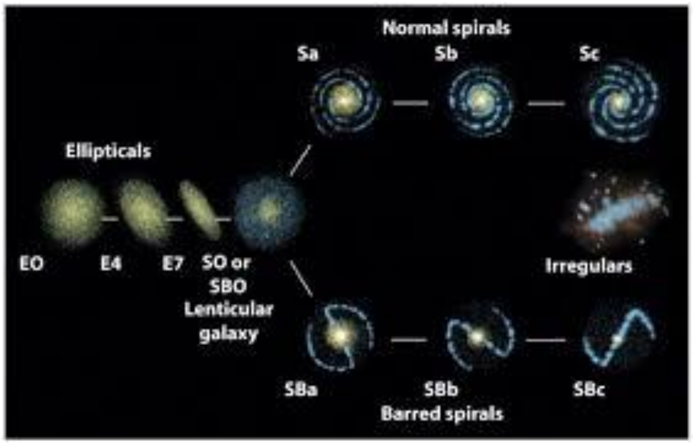


Fig 1: “Hubble Tuning-fork” Classification Technique

II. GALAXY ZOO PROJECT

Human is at on the top of classifying the data of galaxies into these and other structural classes. To that end, this galaxy zoo (GZ) project [4] was launched the classification of a vast sample of SDSS galaxies to the crowdsource. It was a program which could be used by local scientists. Those scientists use it to classify the galaxies by following the programmed decision tree. There are over 100,000 members, who have classified nearly 1 million galaxies by their own hands.[5].Multiple biases had corrected this data many a times and then it is now shown to be in strong agreement with classifications created by professional astronomers.[6].The dataset of this project helped a number of remarkable papers in astronomy.[6-8]. It has also provided priceless training set that perhaps classify billions of more objects with automated classification algorithms. In my project I have used this dataset to train and test different models used in this project. I tried to examine the basic structure of data using unsupervised learning.

III. PROJECT

I have used Galaxy Zoo dataset [9] in my project. In this dataset I have 16 columns and 667945 rows. In these rows first row is of headings of the particular columns. It has last 3 column which give the information about the structure of the galaxy. Fourteenth row tell that the particular galaxy is whether spiral or not. Fifteenth column tells that whether the galaxy is elliptical or not. And the last, sixteenth column tell if the galaxy is of irregular shape. I have added an additional column named "Class". Initially it is an empty column. In the coding this column will be filled according the shape of the galaxy that is if the galaxy id is spiral, the class value become 0, if it is elliptical, then it will become 1 and if it is irregular, then the class value become 2. As the dataset is very large, I am using a small fraction (.01%) of this dataset. This fraction will be done randomly that is the dataset will change every time the code executes. In order to add value to the column "Class" I need a continues index, so I have added a column named "ID" as first column. By using the values of this column, the value of other column like spiral, elliptical and uncertain are accessed and the value of class is filled in the correct the row. After that the columns named "Spiral", "Elliptical" and "Uncertain" are dropped as they are of no use now. Then 2 column named "RA" and "DEC" are also dropped because the values of these column are in string format which can't be used in classifiers to classify the class. Then the dataset is divided into 2 parts as "x" and "y". "x" has all column except the last one and "y" column contains the data of the last column. Then it was split the data into training and testing dataset. Then the classifier "SVM" is used with the "linear" kernel. It is giving approx. 60-63% accuracy and then the program printed the confusion matrix and classification report. Then the other algorithms are applied. After "SVM", "Random Forest" algorithm is used and plotted a graph (Fig 2).

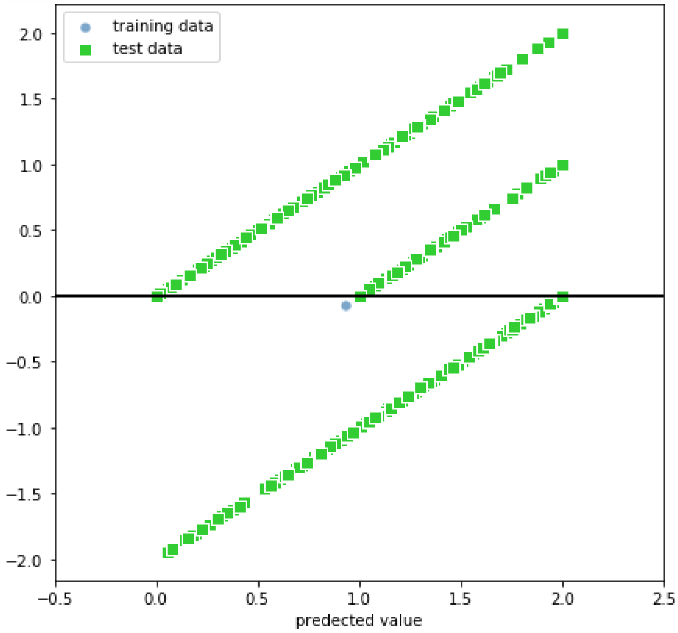
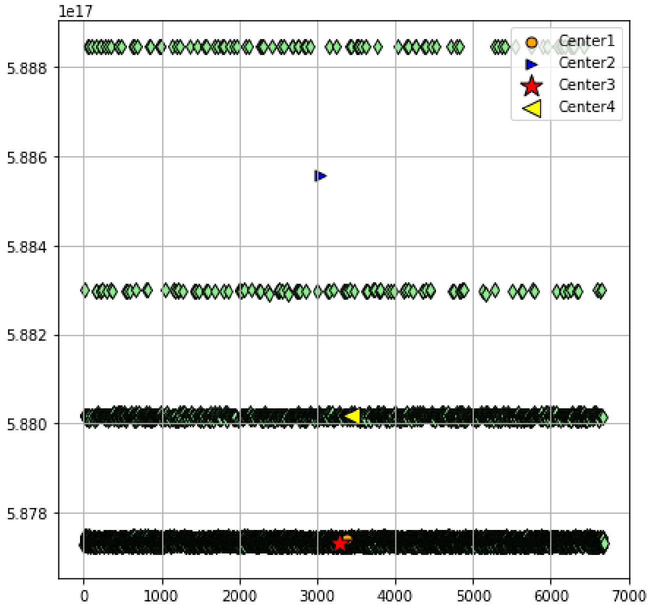


Fig 2: Random Forest Graph

After this algorithm, K-means algorithm is used and its graph is used with its distortion value (Fig 3).



```
Distortion:11801913221780033229092290560.00
```

Fig 3: Clusters

I have used elbow method and silhouette score to find the clusters in this algorithm. Although it depends upon the dataset used, but mostly in the elbow method, the clusters chosen are 3(Fig 4) and with silhouette score method, the number of clusters are 4 (Fig 5).

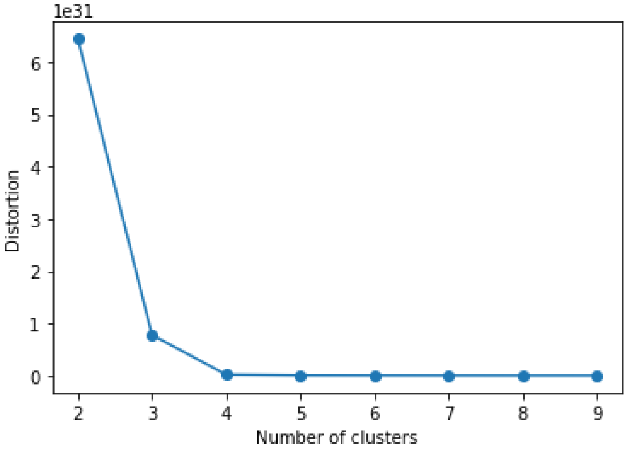


Fig 4: Elbow Method

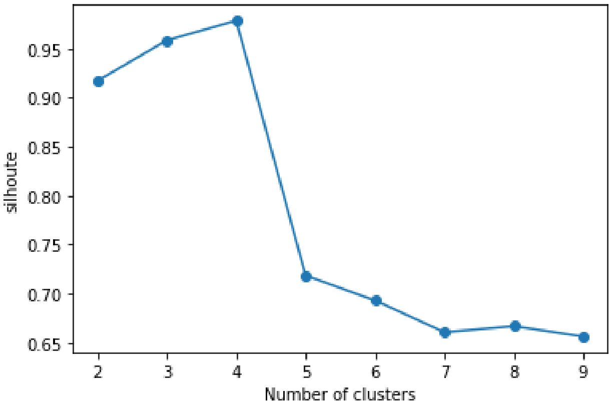


Fig 5: Silhouette Score

After this "Agglomerative Clustering" algorithm had applied on the dataset. It prints the clusters(Fig 6) and after this the dendrograms(Fig 7) were also printed in this program.

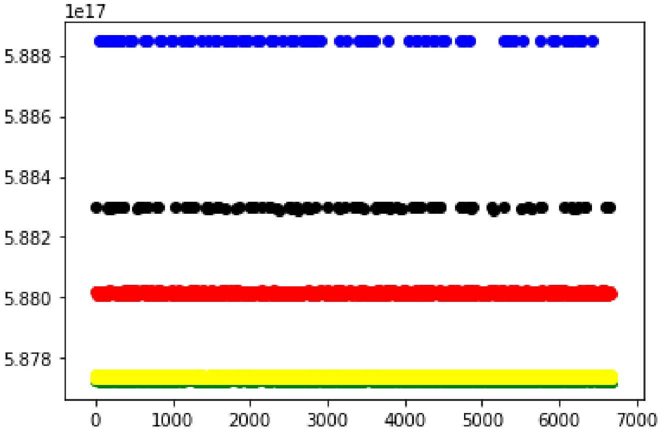


Fig 6: Agglomerative Clustering

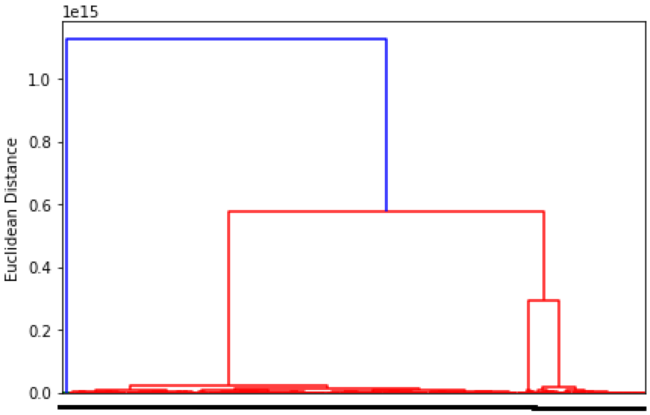


Fig 7: Dendrograms

The last algorithm named "Decision Tree Regression" has applied and then plotted the graph.(Fig 8)

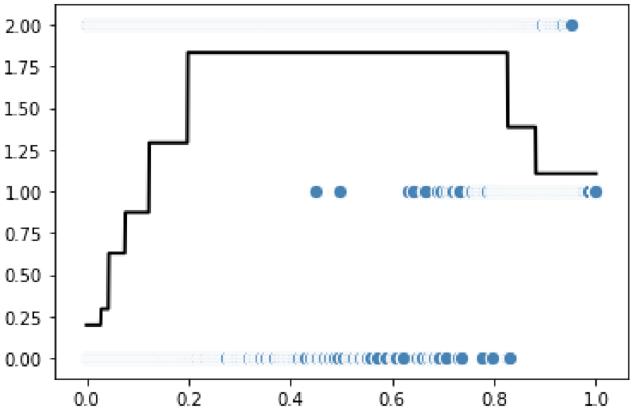


Fig 9: Decision Tree

IV. CONCLUSION:

The main purpose of this paper is to know about the classifications techniques and getting the knowledge of applying these techniques on the dataset of galaxies to classify these galaxies according to their structure. In this project I have used both supervised and unsupervised techniques to classify the galaxies. In supervised, I have applied SVM, Random Forest and Decision Tree algorithm. In unsupervised, I have applied K-Means and Agglomerative Clustering algorithm. With the help of this project the galaxies can be classified into 3 categories: Spiral, Elliptical and irregular (uncertain).

V. REFERENCE:

[1] Galaxy Morphology Classification Paper by Stanford University:

http://cs229.stanford.edu/proj2016/report/GauthierJainNoordeh-GalaxyMorphology-report.pdf

[2] The thirteenth data release of the Sloan digital sky survey:

https://arxiv.org/abs/1608.02013

[3] LSST Science Book, Version 2.0:

https://arxiv.org/abs/0912.0201

[4] Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey:

https://arxiv.org/abs/0804.4483

[5] Galaxy Zoo Wikipedia:

https://en.wikipedia.org/wiki/Galaxy_Zoo

[6] Galaxy Zoo: the dependence of morphology and colour on environment:

https://arxiv.org/abs/0805.2612

[7] Galaxy Zoo: a sample of blue early-type galaxies at low redshift:

https://arxiv.org/abs/0903.3415

[8] Galaxy Zoo: the large-scale spin statistics of spiral

galaxies in the Sloan Digital Sky Survey:

https://arxiv.org/abs/0803.3247

[9] Galaxy Dataset Table 2:

https://data.galaxyzoo.org