Survey of The Sky

Abstract

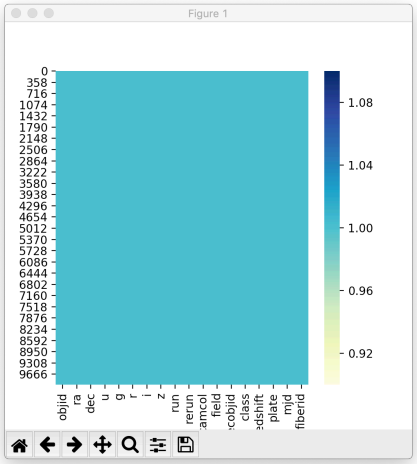
Data science is changing Astronomy. With the rise of big data and super computers, Astronomy is providing large amounts of data from celestial observations. One of those observations is to classify objects in the sky. In my data set, I was given observations made by a multi imaging spectroscopic redshift survey, through which I want to classify different types of celestial objects. The objects to classify are Galaxies, Stars, and Quasars. In classifying, I wanted to determine which features of the data set would be the most important, how I could solve the imbalance between classes among the data set, and lastly what is the optimal method when determining classification through data mining techniques. The results was simply that the redshift feature was the most important, the oversampling of the minority class on my data set proved to increase accuracy, and decision trees evaluated to be the optimal classifier.

Introduction

I had the opportunity to attend a lecture given my Dr. Michelle Thaller, Assistant Director of Science Communication at NASA Goddard Space Flight Center. In her talk about the state of NASA, she mentioned when taking UV (ultra violet) images of the Sun, the data would be so large it would consume 2,600 terabytes per minute. This was the moment I wanted to connect data mining with Astronomy. I was able to find a data set called “Survey of the Sky”, which was a collection of readings from a multispectral imaging and spectroscopic redshift survey. This optical telescope is located in New Mexico, and is meant to create a 3D mapping visualization of our visible universe. The Sloan Digital Sky Survey provided a dataset for observations made by their telescope. This data set included many features, but among them was the class: whether this object in the sky is a Galaxy, Star or Quasar. A galaxy is a collection of stars, and a star is simply a luminous spheroid that is held together by hydrostatic equilibrium and gravity. However, Quasars are a type of nucleus in galaxies, specifically Active Galactic Nucleus (AGN). Simply put, it’s the huge concentration of light in the center of a galaxy. Each data instance contained these three objects, which inspired three main problems for me to solve: How could I solve imbalance for this classification problem, Which classifier would be the optimal one for classification through data mining and why, and lastly which feature has the highest importance factor. The problem of determine which feature has the highest factor is an important problem to solve, and I will take two routes toward solving the question. One will be deterministically, meaning I will find an algorithm to determine this solution, and second will be non deterministically, in that I would develop a solution through research into topics relating to celestial objects.

Solution

My first approach to solving all problems was to understand the data, and this meant plotting aspects of my data such as whether the data set contains null values, or learning the imbalance of of classes.



Heat map of null values for features

Class Distribution

To solve the problem of imbalance in my data set, I researched the best possible ways using a myriad of Medium articles that were in the Data Science channel. The recommended method for solving imbalance issues in data was to use intelligent metrics, and to use the idea of under sampling or oversampling. For using intelligent metrics, I couldn't simply use accuracy to determine if my classification results were good. I used the F1 score which is the 2 * ((precision * recall) / (precision + recall)). This metric is important as it takes into account the recall which is needed due to the imbalance nature of my data set. Another technique used to help with my imbalanced data set is oversampling the minority class. Under-sampling wouldn't work as well because my data was small in size, and partitioning this data set into training and validation makes it even smaller. I would in this case want more data, hence why oversampling was chosen. To accomplish this, I used the SMOTE (Synthetic Minority Oversampling Technique) algorithm, which is an algorithm for taking a set of data containing the minority class, and synthesizing more data sets containing that minority class. To solve the problem of determining what is the best method to create a classifier for this data set, I chose classifiers that are simple in understanding. I chose KNN (K-Nearest Neighbors), NN-MPL (Neural Network Multi layered Perceptron), Naive Bayes, Linear SVC, Decision Tree, and lastly Random Forest Tree. All classifiers were trained and fitted using Sci-Kit learn libraries. I would proceed by then obtaining the F1 score using the metrics library provided by Sci-Kit learn, and plotting a bar graph of each classifier with it's respective F1 score. For the final problem of determining which feature has the highest importance, I took two different routes as mentioned earlier. For

determining the feature with the highest importance deterministically, I used a technique called “model feature importances”. This is a property given from the classifier of a Random Forest Tree that allows you to view the various attributes of your data, and determines which one is the most important via a double number. It does this by calculating the GINI Importance which calculates each feature as a sum over the number of splits across the entire tree that includes that feature, divided by the number of samples it splits. For my non-deterministic route, I chose to read a text book called “OpenStax Astronomy” which is an introductory Astronomy Text book. Chapter 6 “Astronomical Instruments” and Chapter 17 “Analyzing Starlight” helped explain concepts such as Redshift, optical filters, light, wavelength and almost everything one needs to know to understand how important light is towards detecting celestial objects. Through the reading of this text book, I was easily able to learn which feature had the highest importance when it comes to classifying the three different types of celestial objects.

Experiments

Data

The data received from the Sloan Digital Sky Survey Website was a CSV file containing 10,000 instances of data, each having 17 features. The data was an observation of different celestial objects in the sky. Some of the features were Redshift, filters of different lights (such as red light, blue light etc.), and location. Location was difficult in understanding, as on Earth to locate something, we use Latitude and Longitude as our coordinates, but what about the sky? I learned we used Right Ascension and Declination as our respective “Coordinates to our sky”. We do this by taking what we currently see in the sky, and mapping this onto a spherical projection called the Celestial Sphere. It is in this Celestial Sphere where we locate this data using Right Ascension and Declination. Understanding the data was an important task in my project, as many of these features were new to me. I had to understand features named “u, g, r, i, z”. This is not intuitive, which required a lot of work to understand. In doing so, I found that these letters are simply filters the telescope uses to receive a certain magnitude or wavelength of light such as Red, Bluish Yellow, Ultra violet etc. For the features of “run, rerun, camcol, field”, they were simply identifying what part of the telescope was involved in a specific data entry. “Plate, mjd, fiberid” are respectively the positioning of optical fibers that is used to direct light, the date on which the data was taken, and the associated fiber used with each data entry.

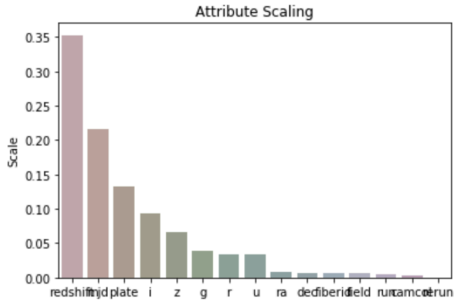
Experimental Setup

For the data, I simply read in a CSV file, and converted it to a numPy array, using the numpy library. My first job was in sanitizing the data, so making sure there were no null values, ensuring there was no missing data, and both of these checks turned out to not be needed. Next, when testing different classifiers I split my training data into 80% training and 20% testing. These were arbitrary values chosen of what the norm usually is. When splitting the data, one key factor I ensured is the sampling of data was not truly random. I needed this imbalance factor to stay the same in the testing and training. To do this, I use a technique called stratified imbalance to provide equal skewness in the testing and training data for all classes. I also reduced the number of features using the deterministic method discussed earlier. I was able to remove 10 features, as results showed 7 of the features had a GINI Impurity score that was insignificant. The performance metrics I chose to use is the F1 score, mainly because it takes into account recall as well as precision. For my technical setup, I used Python 3.7 as my language of choice. I used Anaconda as a Data Science platform for me to use Jupyter Notebook for fast implementation of my solutions, and Sci-Kit Learn as my library dependency for obtaining various classifiers. The seaborn library was also used to easily

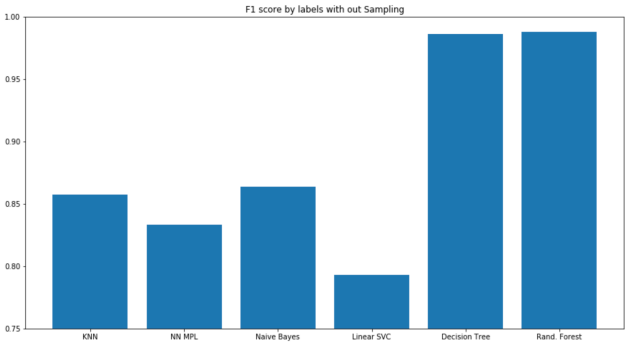
visualize data using bar plots, heat maps, density charts and many other types of visualizations.

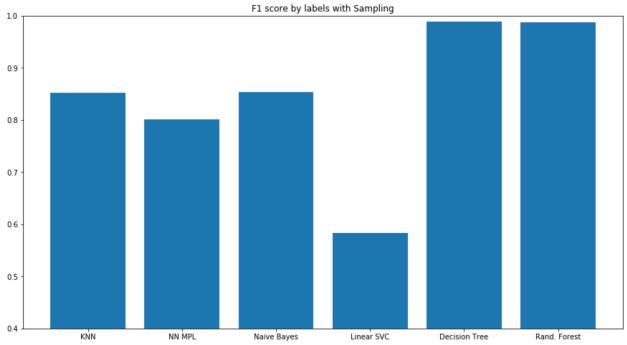
Experimental Results

For determining which feature would be the most important in classification, I concluded Redshift. When reading the textbook, I understood that redshift was the light collected by the telescope and being analyzed to determine whether an object was moving closer or farther away from us. Scientists did this by looking at the wavelength of light wave. The equation for redshift is *"Redshift = (Observed wavelength – Rest wavelength)/(Rest wavelength)"*. If the wavelength is getting larger, then it is shifting towards the red spectrum (the longer wavelength spectrum) and it is inevitably getting farther away from us. This is on par with the theory that the universe is expanding. OpenStax Astronomy also stated "One way quasars had to obey the Hubble law was to demonstrate that they were actually part of galaxies, and that their redshift was the same as the galaxy that hosted them…". This allowed me to deduce that the Redshift actually allows one to see the distance of a celestial object. The farther the object, the larger the redshift. And this was key for classification. Telescopes cannot make out a star within another galaxy, so stars will always be closer to us than galaxies, hence stars' Redshift will be smaller in magnitude. This made sense in the data, as the average magnitude of Redshift for stars was around 0.00, and for Galaxies and Quasars it was ~0.2. This made it easy to differentiate Quasars and Galaxies as they will both have similar redshift values, and will be able to differentiate based on the light only. Quasars will emit different light than galaxies as galaxies are composed of all different types of stars, however a Quasar is simply composed of its self. Deterministically, this was correct, as the Random Forest Tree's attribute on feature importances showed that redshift was the most important feature.



The results for the problem of solving imbalance in my class was straightforward, as oversampling using the SMOTE algorithm would help with this imbalance. In the results relating to which classifier would have the best performance via F1 score, I performed two different tests: using oversampling and not using oversampling. The results were very minimally different. However, the Decision Tree classifier was the best in classifying celestial objects. My reason as to why is because decision tree's and random forest tree's are applying cost sensitive learning which applies a larger misclassification weight to classes that are skewed. This is why I believe the tree's performed the best. The only significant change through resampling was the SVC classifier. It when from a 0.79 F1 score to 0.58. The rest relatively was +-(0.02).





Classifiers	F1 Score
KNN	0.85
Neural Network (MLP)	0.83
Naive Bayes	0.86
SVC	0.79
Decision Tree	0.98
Random Forest	0.98

Classifiers	F1 Score
KNN	0.85
Neural Network (MLP)	0.83
Naive Bayes	0.85
SVC	0.58
Decision Tree	0.98
Random Forest	0.98

Conclusion

Data Mining combined with Astronomy can bring about a powerful set of tools. Currently, this dataset is classified through images. There is a number issues to this as images take up a lot of data, and classifying images visually takes up a lot of time. These commodities of data and time can be saved through the work of data mining and creating classifiers such as the one presented in this paper. Overall, this paper shows that redshift is the most important feature when it comes to classifying celestial objects, decision tree's are optimal for obtaining the best F1 score for classifiers in this dataset, and oversampling helps increase the F1 score and is a valid way to classify against this imbalanced dataset.

Contribution

I, Sahil Gangele, performed this project alone. I contributed 100% towards the completion of this project.