Data Science – Classification of Large Astronomical Photometric Datasets

Project Description

Astronomical datasets are typically very large. Processing these datasets (such as the dataset available at the Sloan Digital Sky Survey (SDSS), even with the most efficient algorithms, takes enormous amounts of time and resources. One of the problems faced while processing astronomical data is to accurately and efficiently classify an object recorded through a telescope.

SDSS cannot get spectrum data for all its objects as it takes about an hour to measure each spectrum and to get data for all the objects viewed would take hundreds of years. There is a total of 1,183,850,913 objects in the Photometric Catalogue, out of which only 3,751,358 are spectroscopically classified, the others are unclassified data.

In this work, we will design the model that can reliably and efficiently identify quasars in large unclassified datasets. We use the datasets available at the Sloan Digital Sky Survey (SDSS). Our model will be trained on the photometric data of 500,000 objects already labelled and spectroscopically classified, which is available from the SDSS spectroscopic data catalogue. This model then uses only the photometric data from the 1.2 billion unclassified objects in the SDSS photometric catalogue to classify them into quasars, stars, etc.

Data Description

In the dataset, available at SDSS, the data which is of interest to us, are of two kinds—spectroscopic and photometric. The objects in the Spectroscopic Catalogue each have a well described spectrum. Based on the spectral characteristics like redshift, emission peaks, absorption peaks, etc., each object has been classified as a star, quasar, galaxy, etc.

The Photometric Catalogue contains color data about all the objects viewed so far, including the objects in the Spectroscopic Catalogue. The color of the object is measured in 5 filters: ultraviolet (u), green (g), red (r), and infrared (i and z).

Project Approach

We will integrate k-Nearest Neighbors (kNN) into Support Vector Machine (SVM) which can run on Spark HDInsight and create an ensemble algorithm of these. It will be used to efficiently identify quasars in a massive dataset of unclassified, photometrically defined objects. Our model's performance will then be cross-validated and its efficiency is estimated on the large dataset. If there is scope of increasing efficiency of our model using other algorithms, we will incorporate them into the existing model.

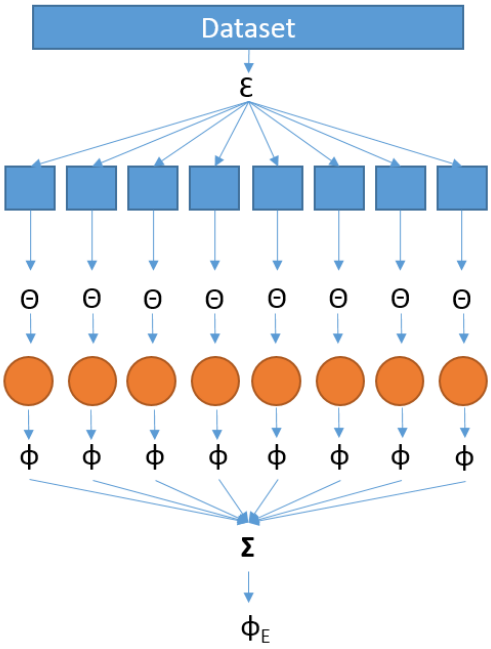

Figure 1 A visual representation of the workflow of ensemble learning and ensemble classification. Initially the dataset is divided into subsets using some subsampling technique ε. Each subset is used to train a classifier using some training scheme Θ. Each classifier then makes individual predictions Φ on input data. The predictions are combined using some voting scheme Σ to produce a final prediction Φ E

Project Architecture

We will use Spark cluster in HDInsight for running our machine learning algorithms. We will use one A4 series instance as master node for the cluster and 7 A4 series nodes as worker nodes for the cluster. We will be running our machine learning algorithms in master node which in turn distributes its jobs among all the worker nodes. The results from all the worker nodes is then collected and is stored over the cloud storage as a text format file

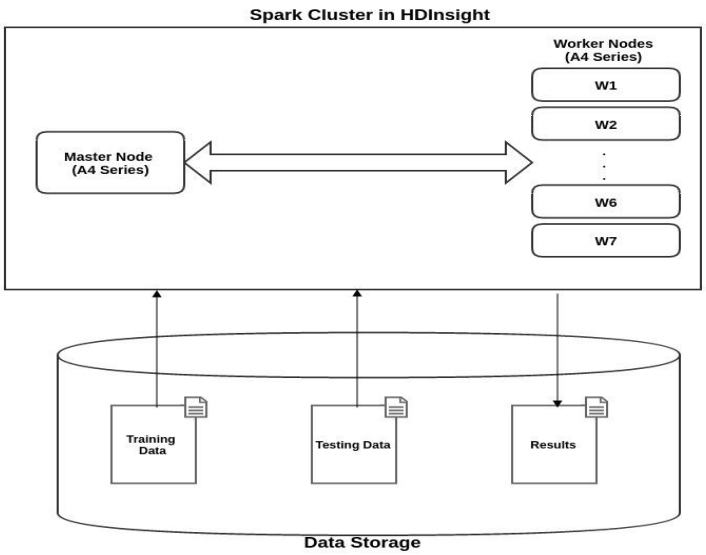

Figure 2 Architectural Diagram

Project Impact

Classifying 500,000 objects spectroscopically would have taken approximately 57 years using astronomical tools, but using our predictive models this can be done in few minutes for the same number of objects. The 1.2 billion unclassified objects available in the SDSS data would therefore take very less time compared to what would have taken to classify those objects using astronomical tools. It can typically be done in about few hours using our model. Therefore, this model is potentially of enormous use to the astronomy community.

Project Plan

1. The Master Node reads the Astronomical Training Data from the storage
2. Using Machine Learning algorithms(like kNN, SVM and ensemble etc), we design a model based on the g-r, u-r, r-i, i-z parameters of the data
3. By using the designed model, we predict the classes of the 1.2 billion unclassified objects over the distributed cloud environment using Spark framework
4. All the predicted classes with their object ID's are then stored in a text file into the cloud storage

Project Requirements

- HDInsight for spark framework
- Eight HDInsight A4 instances of 8 cores and 14GB RAM. One for the master node and seven for the slave nodes.
- Storage of 1TB
- Each node requires 1000 hours of compute time. So, totally 8000 hours for 8 nodes.

References

1. http://www.sdss.org/dr12/
2. D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. Allende Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot, and et al., "SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems," The Astronomical Journal, vol. 142, p. 72, Sep. 2011.