Clustering of Cross-Referenced Astronomical Data Sets

Vera Abaimova — stormseecker@gmail.com
Mark Wells — mwellsa@gmail.com

December 9, 2014

Abstract

The large amount of astronomical data that is now available from a multitude of missions creates a need for machine learning methods that can analyze it and glean information about our universe. One challenge in particular is to analyze data about the same observations made by different missions, *i.e.*, analyze cross-referenced data, especially data in different formats. One such format is time series data provided by the *Kepler* mission, which results in additional difficulties. Our proposed method extracts features from cross-referenced data sets, including time series data, and applies a hierarchical clustering model in order to aid stellar classification.

1 Introduction

1.1 Motivation

Astronomy and astrophysics is a constantly growing field that is heavily data-dependent. In the near future, petabytes of data will be procured on any given night by the various telescopes and scientific missions tasked with surveying the night sky in the search for objects of interest. Such a large quantity of data means that there is a dire need for quick object identification and classification in order to obtain useful knowledge from the data, a task that has quickly grown beyond the ability of unaided humans.

The introduction of machine learning into the field of astrophysics has resulted in a new discipline that combines the two areas of research: astroinformatics. The goal of astroinformatics is to harness the power of machine learning and data mining in order to find new and interesting knowledge from the stockpile of data that is constantly being produced at an accelerated rate.

The numerous astronomical surveys currently in operation often times (and not by accident) generate data on overlapping parts of the sky. Each mission generates different types of information: spectral, spatial, temporal, or any combination. The result is that some of the same stars will appear in multiple data sets and will have more diverse data available via the contribution of the different missions. Unfortunately, not a lot has been done with this synergy, as most of the current literature indicates that the majority of effort goes to classifying stars using one data set alone. However, there are a few efforts out there that maximize their study by combining data from multiple (and varied) data sets.

Ultimately, the acquisition of data and its interpretation into interesting information will help humanity better understand the universe in which we live.

1.2 Challenges

There are numerous challenges that arise when trying to correlate different astronomical data sets. Routinely, the data is in different forms and of different resolutions. Sometimes there is target confusion which makes it difficult to separate the received signal from two or more potential sources. For example, there could exist a background star and its variability could be falsely detected when measuring the foreground star of interest.

Other factors also serve to compound this already difficult process. Missing data is very common: the spacecraft had to slew to a transient event, the post-processing performed by the mission team had to truncate data, or the object was too faint to be observed. Since these factors are spacecraft dependent it is quite possible that even when two stars are in both data sets, pertinent data could be missing from one of the sets which would exclude its utility.

When it comes to analyzing the data it is sometimes difficult to correlate the separate features in a useful way. Obviously the location of an object in the sky seldom has any relevance (save in certain unique cases). Also, bias could be an issue. Stars, because of their inherent properties, could be excluded from one dataset but not another, so entire classes of objects are underrepresented. With analysis of time series data, the sheer size of the data presents challenges. It is difficult to work directly with all of the data points in such a format. How we mitigated this problem is further discussed in Section 4.

Finally, the sheer amount of data available and even the data set that we ended up working with is incredibly large so it took a long time to process and it was computationally expensive to analyze.

1.3 Related work overview

In the existing literature there are several data mining approaches that are employed when analyzing astrophysical and astronomical data. These fall into the following camps: supervised learning, unsupervised learning, and semi-supervised learning. We will only be concerning ourselves with unsupervised learning, specifically clustering, and we will also take a look at distributed data mining.

Clustering is an extremely popular unsupervised method in the field of astrophysics. Distributed data mining is an approach that is useful for large data sets, among other qualities, so it will be reviewed as well. All of these methods will be discussed in further detail as part of the related work in current literature.

1.4 Proposed Method

Our proposed method consists of cross referencing three data sets: *Kepler*, *GALEX*, and *SDSS* and then extracting the relevant features. The *Kepler* data set also provides time series data, so part of the feature acquisition process involved processing the time series lightcurves and extracting useful features from them. Before we could do that though, we performed a process called flattening, which we did twice, once over each quarter and again over each day of data. After the flattening process we extracted the variance, skewness, and kurtosis of each lightcurve and included the resulting values in our feature space.

After the data preprocessing we performed an analysis. Our chosen baseline method is hierarchical clustering, which perfomed better than the K-Means clustering we performed for our preliminary data analysis. More in depth explanations of both our methods and experimental results are further outlined in Section 4 and Section 5, respectively.

1.5 Overview

In Section 2 we will give a short survey of the literature which outlines existing approaches to processing time series data and working with large data sets. The next section, Section 3 will a formal, mathematical definition of our problem, including the input and the expected output. Section 4 is where we will go more in depth, first about the baseline methods we tried in our preliminary analysis of the data, and then about our final data analysis. Section 5 will contain the information about our data sets, our experimental results, both for our preliminary and final passes, and an analysis of our final results. The final section is a summary of the problem and our experimental results, along with what we have learned from this project.

2 Related Work

3 Problem Definition

The input of our clustering problem is a list of $x_{i,j}$ where x_i is the star object, numbered from 1 to 20,840 and j is the feature, numbered from 1 to 29. Our feature space includes features such as fuv_mag, nuv_flux, and *kepmag*, among others, where fuv_mag is the far ultraviolet magnitude, nuv_flux is the near ultraviolet, and *kepmag* is a measure of brightness of an object in the *Kepler* pass band.

Our output is a list of $x_{i,j} \in C_k$ where C is cluster and k denotes cluster number.

4 Method

5 Experiments

6 Summary

References