Detecting variability in astronomical time series data: applications of clustering methods in cloud computing environments

Min - Ju Shin1, Young - Ik Byun2, Seo - Won Chang3, Dae - Won Kim1, Myung - Jin Kim4
Dong - Wook Lee5, Jangyoon Han6, Yong - Heung Jung7, Jeannon Youn8, Jee - Hyuk Kowalc9, Joo Hyun Kim9
1 University of Michigan, 2 Yonsei University, 3 CfA, 4 KASI, 5 KISTI

We present applications of clustering methods to detect variability in massive astronomical time series data. Focusing on variability of bright stars, we use clustering methods to separate possible variable objects from other time series data, which include intrinsically non-variable sources and data with common systematic patterns. We already achieved the analysis of the Northern Sky Variability Survey (NSVS) data, which include about 16 million light curves, and present candidate variable sources with their association to other data at different wavelengths. We also apply our clustering method to the light curves of bright objects in the SuperWASP Data Release 1 (DR1). This study is conducted in cloud computing environment provided by the KISTI Supercomputing Center. We explain our experience of using the cloud computing test bed.

1. Introduction

Detecting variability in astronomical time series data can be understood as a problem of outlier or anomaly detection in statistics and machine learning. The main assumption is that the time series data of detectable variable objects is considerably different from the rest of data. In usual data sets of astronomical time series data, there are considerably more normal objects, i.e., non-variable objects, than abnormal objects, i.e., variable objects. When we can describe the data properties of non-variable objects well in given data sets, we can detect variable objects easily and completely. The common types of anomaly detection methods are graphical and statistical-based, proximity/distance-based, density-based, and clustering-based. Traditionally, astronomers have used statistical-based abnormal detection methods, while generally ignoring any systematic patterns of time-series data or erasing the systematic patterns with empirical approaches. Here, we apply clustering methods, which can consider objects affected by common systematic patterns as ordinary types, and can separate peculiar objects as variable candidates.

2. Methods: clustering with variability indices

Multiple variability indices

In order to detect various patterns of light variation, we adopt the following variability indices which are derived from the SuperWASP DR1 does not have well-defined observation epochs, more than 15 epochs, as variable and non-variable candidates in 638 NSVS fields.

3. Variable candidates in the NSVS

We cluster 16,189,040 light curves, having data points at more than 15 epochs, as variable and non-variable candidates in 638 NSVS fields.

4. Processing SuperWASP DR1

We also explore the public data release 1 from the SuperWASP project. This data set has about 15 million light curves over both northern and southern skies. Because the SuperWASP DR1 does not have well-defined observation fields, first, we group light curves having similar time-scales such as starting and ending times.

Cloud computing environments

For processing the SuperWASP DR1 data, we try to use a Cloud computing testbed which is deployed by the Korea Institute of Science and Technology Information (KISTI) Supercomputing Center. We test two different system configurations. One system uses Condor as a job management software, and stores data in Lustre distributed file system. Another system adopts Hadoop computing environment with its distributed file system. Both systems are built with virtual machines which are managed by Eucalyptus. Although Hadoop systems does not allow us to use different file systems and is less flexible than the Condor system, its job management considers data locality which can improve performance of parallel distributed processing.

5. Conclusion

We are processing all SuperWASP DR1 light curves to find new variable candidates by using the KISTI Cloud computing environment. The most parts of the processing will be done with the Condor system. The analysis results of the NSVS and SuperWASP DR1 will be released to public on our web site (http://stardb.yonsei.ac.kr) soon.

Infinite Gaussian Mixture Model

Gaussian Mixture Model (GMM) is commonly used as a density estimator and a clustering method by describing the distribution of data as mixtures of multivariate Gaussian distributions. Because we do not know how many clusters exist in our data, we use an Infinite Gaussian Mixture Model with Dirichlet Process which allows the model to have infinitely many mixture components:

\[ p(x) = \sum_k \pi_k \frac{1}{\sqrt{2\pi} \sigma_k} \exp \left( -\frac{(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)}{2} \right) \]

where \( n \) is an index over \( M \), \( x \) is an \( N \)-dim vector of parameters, and \( \gamma \) is the number of parameters (in our case \( \gamma = 8 \)). The final distribution of all objects is given by:

\[ p(x) = \sum_k \pi_k \gamma_k \frac{1}{\sqrt{2\pi} \sigma_k} \exp \left( -\frac{(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)}{2} \right) \]

where \( \gamma_k \) is the fraction of each mixture component.