

Detection of Rare Objects in Massive Astronomical Datasets Using Innovative Knowledge Discovery Technology

A. Soto, A. Cansado, and F. Zavala

Department of Computer Science, Pontificia Universidad Catolica de Chile, Santiago 22, Email: asoto@ing.puc.cl

Abstract. Our work presents an application of knowledge discovery technology aimed to help scientists in the detection of rare types of astrophysical objects. Our main idea is that while computer have the power to search huge amounts of data, an expert has the domain knowledge to efficiently lead this search. Our system builds upon two main components: a probabilistic model able to scale to large datasets and a set of modules to interact with the scientist. Here, we focus on the probabilistic model used to represent the joint uncertainty among the attributes of the objects registered in a sky survey catalog. This model consists of a combination of a Bayesian network and a set of Gaussian mixture models (GMMs) trained with an accelerated version of the expectation maximization (EM) algorithm. The model is currently being tested using data from the release 1 of the Sloan Digital Sky Survey. The results indicate that the system is able to accurately detect a set of simulated rare objects, but it also provides a large number of false positives.

1. Introduction

Today's Astronomy is living an information revolution, critically needing novel technologies that help in the analysis of new vast sources of data. Several recent projects to massively survey the sky are starting to generate many Terabytes of data, with billions of sources detected and tens or hundreds of parameters measured for each of them. Up to date, researchers have mainly handled these types of surveys manually. However, this approach is becoming no longer viable. As Astronomy is expanding its frontiers, automated knowledge discovery in databases (KDD) is emerging as a key technology to take advantage of the new data available.

There are several interesting applications where KDD technology may increase the efficiency of the analysis of massive astronomical datasets, such as cataloging the information of huge amounts of sky images, automating the clustering of galaxies, and developing better interactive visualization tools, among others. In particular, one of the most prominent new applications is the detection of unusual objects.

The discovery of rare or new types of astrophysical objects plays a key role in Astronomy (red shift quasar, L and T starts, brown dwarfs, etc.) (Djorgovski

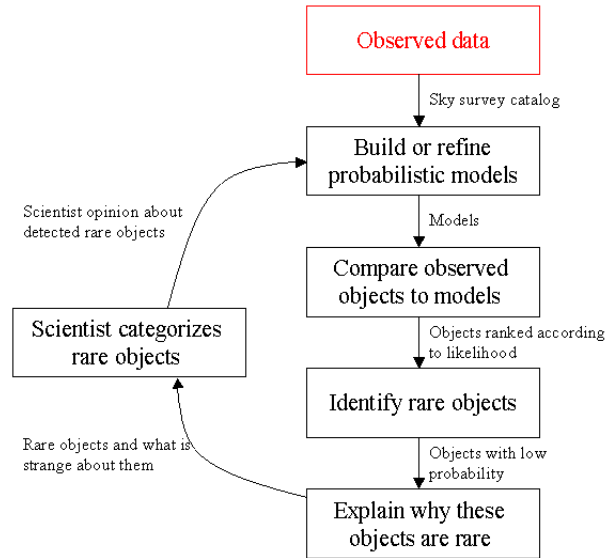


Figure 1. Diagram of the intended application.

et al. 2001). With the next generation of massive sky surveys, the search for new phenomena in our Cosmos will intensify in the close future. There will be wider frequency ranges available, and significantly more sources and resolution, increasing the chance of discovering and studying new or unusual objects.

In terms of KDD technology, the detection of unusual objects in a sky survey catalog is understood as the detection of outliers in a database. Several KDD techniques can be used to solve this problem (Kou et al. 2004), however, their application to high dimensional and massive datasets faces devastating computational and statistical difficulties. Problems such as the curse of dimensionality, lack of labeled data, and an extremely low number of outliers prevent the use of off-the-shelf technology.

In this work, we propose the creation of a computing tool to aid the scientist in the discovery of rare types of astronomical objects. While computers have the power to search huge amounts of data, an expert has the domain knowledge to lead the search. This task requires the development of new algorithms to robustly identify anomalies in huge data sources. It also requires the presence of active learning, to intelligently decide which record to display to the human scientist. Both of these steps must be scalable to interactive speed on billion-record data.

Figure 1 sketches the main features of our approach. Using observations from a sky survey catalog the system is able to learn a probabilistic model of the objects in the database. Using the model, the system sorts the objects in the catalog according to their likelihood. Highly common objects like typical stars or galaxies will show a high likelihood, but rare objects will be poorly explained by the model, thus, they will show a low likelihood.

Using the previous intuition the system selects the low likelihood points as a set of possible rare objects. Next, the system analyzes the attributes of the

rare objects aiming to find an explanation about what is strange about them. Finally, the system sequentially presents to the scientist the most prominent rare objects, together with an explanation of what makes them unusual. At the same time the system uses the scientist feedback to interactively improve its probabilistic model for the detection of rare objects. This online improvement of the model is a form of active learning that allows an efficient use of the scientist feedback, helping to focus the efforts on key steps of the algorithm and avoiding to reprocess irrelevant objects.

In this paper we focus on describing the main steps of the method used to build the joint probability distribution that models the data. This model consists of a combination of a Bayesian network (Pearl 1988) and a set of GMMs (Duda & Hart 1973) trained with an accelerated version of EM algorithm (Dempster et al. 1975).

2. Probabilistic Model

Our probabilistic model considers the joint probability distribution (JPD) of the attributes of the objects in the sky survey catalog. Using a Bayesian net, we take advantage of conditional independence relations among the attributes to obtain an efficient factorization and graphical representation of the JPD. These independence relations divide the JPD in simpler local conditional probabilistic distributions whose reduced dimensionality simplifies the estimation process.

The novelty of our approach resides in the incorporation of special techniques that provide the efficiency required to scale the approach to huge sky survey catalogs. This is especially critical since a slow KDD analysis severely limits the interaction between the scientist and the data. The probabilistic model provides 3 main features to our algorithm: scalability to large datasets, efficient detection of rare objects, and explanation of what makes an object unusual. As for scalability and detection, Sections 2.1. and 2.2. describe the 2 main steps involved in finding an appropriate Bayesian net: learning the structure of the network and learning the conditional probability distributions that relate the nodes in the network. Next Section 2.3. explains our approach to find what makes an object unusual.

2.1. Learning the structure of the Bayesian Network

To learn the structure of the network, we use a variant of the Sparse Candidate algorithm (Friedman et al. 1999) that shares the same general basics steps: Restrict and Maximize. The first selects candidate parents for each variable, and the latter uses Greedy Hill Climbing and BIC scoring criteria to find the network structure with maximum score according to the current constraints. GMMs trained with EM are used to model continuous variables and to sample from the net. All calls to EM algorithm are cached for later use. In the estimation of mutual information in the Restrict-step the variables are discretized using an adaptive approach. Further details of the algorithm are given in a forthcoming paper.

2.2. Learning the local conditional probability distributions

To learn the local conditional probability distributions, we train the GMMs with an accelerated version of the EM algorithm. This version is based on the use of condensed representations (Moore, 1999). These representations exploit the internal operation of EM to pre-computed statistical information that summarizes parts of the database. The intuition is to replace a set of similar points by a representative point (Bradley et al. 1998). In this way, at each iteration the EM algorithm does not need to visit each data-point, but only isolated points and the statistics of the clustered points. At each iteration, we cluster or condense points using the membership factors between the points and the gaussians that form each GMM.

One of the limitations of training the GMMs with EM is the selection of the number of components that form the mixture. We tackle this problem using a similar approach to the KD-Clust algorithm (Sand & Moore 2001), however, we do not use KD-trees because of their limitations to operate in high dimensional datasets. We also extend our algorithm to search for convergence points using operators to add, merge, and delete clusters.

2.3. Finding an explanation of what makes an object unusual

One advantage of using Bayes nets is the straightforward evaluation of the likelihood of each data point, which otherwise can be the bottleneck for the detection of rare objects. Furthermore, the local structure and conditional probability distributions embedded in the net provide key information about which groups of attributes are related with other groups, to what extent they are related, and under what circumstances. In this sense, the relative values of the local conditional probability distributions for a given object provide key information to detect the attributes that make the object unusual. We exploit this fact by comparing the attributes of each of the rare objects with the attributes of the most typical clusters in the data. These most typical clusters are given by local relations that explain a large part of the probability distribution.

3. Results

We test the performance of an initial version of the probabilistic model using a subset of data from the release 1 of the Sloan Digital Sky Survey (MySkyServer from SDSS DR1). We first eliminate a small subset of records with missing values, end up with a database consisting of 166.000 records and 194 attributes for each of them. We treat these attributes as continuous variables, except in the calculation of the mutual information, where we discretize the variables to 4 labels. To learn the structure of the network we limit to 5 the maximum number of parents for each node.

In order to test the performance of the model to detect rare objects, we insert in the database 1000 simulated rare objects with random values in some of their attributes. The results indicate that the system successfully identifies all the inserted points as anomalies. The test also classify as rare 10% of the objects in the original database.

4. Conclusions and Future Work

We present an algorithm to aid the scientist in the detection of unusual objects in large databases generated by new massive sky surveys. Considering the simulated objects inserted in the database, the results indicate good capabilities to detect rare objects. However, the 10% of objects considered as unusual in the original database suggests that the system detects a large number of false positives, since according to experts the number of unusual objects in the original database should be less than 0.1%. We are currently working on fine tuning the numerical precision of the probabilistic model, since it is unable to handle the low values associated with the likelihood of most of the false positives. We are also working on including the feedback from the scientist which will be a key factor to focus the search on the real unusual objects.

On other hand, we are testing a modified version of the distance-weighted K-nearest neighbor algorithm to explore the set of rare objects with the goal of optimizing the use of time and feedback from the scientist. We are also modifying the GMM-EM based algorithm, in order to accommodate data with measurement errors. We expect to test the final system on the complete dataset from SDSS DR3.

Acknowledgments. This work is partially funded by FONDECYT grant 1030336.

References

- Bradley, P., Fayyad, U., & Reina, C. 1998, MSR-TR-98-35, Microsoft research
- Friedman, N., Nachman, I., & Peer, D. 1999, UAI, 206-215
- Moore, A. 1999, in Advances in Neural Information Processing Systems 11
- Dempster, A., Laird, N., & Rubin, D. 1977, Journal of the Royal Society, B 39, 1-39
- York, D. G. et al. 2000, AJ, 120, 1579
- Djorgovski, S., Carvalho, R., Odewahn, S. 2001, in ASP conference series, vol. 255
- Kou, Y., Lu, C.T., Sirwongwattana, S., & Huang, Y.P. 2004, in Proc. of Int. Conf. on Networking, Sensing, and Control, 749-754
- Pearl, J. 1988, Morgan Kaufmann
- Duda, R., Hart, P. 1973, John Wiley and Sons
- Sand, P., Moore, A. 2001, in Proc. of Int. Conf. on Machine Learning,