Highlights of Spanish Astrophysics VI, Proceedings of the IX Scientific Meeting of the Spanish Astronomical Society held on September 13 - 17, 2010, in Madrid, Spain. M. R. Zapatero Osorio et al. (eds.)

Data mining in the Spanish Virtual Observatory. Applications to Corot and Gaia

Mauro López del Fresno[1], **Enrique Solano Márquez**[1], **and Luis Manuel Sarro Baro**[2]

[1] Spanish VO, CAB (INTA-CSIC). P.O. Box 78, 28691 Villanueva de la Cañada, Madrid Spain
[2] Departamento de Inteligencia Artificial. ETSI Informática. UNED. Spain

Abstract

Manual methods for handling data are impractical for modern space missions due to the huge amount of data they provide to the scientific community. Data mining, understood as a set of methods and algorithms that allows us to recover automatically non trivial knowledge from datasets, are required. Gaia and Corot are just a two examples of actual missions that benefits the use of data mining. In this article we present a brief summary of some data mining methods and the main results obtained for Corot, as well as a description of the future variable star classification system that it is being developed for the Gaia mission.

1 Introduction

Data in Astronomy is growing almost exponentially. Whereas projects like VISTA are providing more than 100 terabytes of data per year, future initiatives like LSST (to be operative in 2014) and SKY (foreseen for 2024) will reach the petabyte level. It is, thus, impossible a manual approach to process the data returned by these surveys.

It is impossible a manual approach to process the data returned by these surveys. One of the most important objectives for the Virtual Observatory (VO)[1] initiative, is to provide access to astronomical databases in an easy, friendly and efficient way. It also aims to provide tools and processed data to the scientific community. Although the work done by itself can not be taken lightly, the VO can benefit greatly by using data mining techniques.

The Spanish Virtual Observatory (SVO)[2] has been working for several years in the adoption of data mining tools into VO services. Corot has already been favored because of

[1] www.ivoa.net
[2] svo.cab.inta-csic.es

this approach, providing automatic stellar variability classes which were not included in the initial list of Corot products.

The Gaia mission, which will gather light curves for around 10^9 stars, has incorporated right from the beginning the necessity of using data mining techniques for returning value added information. The SVO has also a relevant presence in the development of a variable star classification system for this mission.

2 Data mining techniques

2.1 Supervised classification

Supervised classification consists in training a system in order to return class assignments for new instances. In the training set we define what classes we are trying to identify on the basis of a set of attributes usually obtained from other surveys. For that reason, the process is also known as experience base learning. Training sets are usually very small in comparison with the number of instances to be classified, and the classification of new instances is fairly quick.

Neural networks [5], support vector machines [7], decision trees [10] and bayesian networks [6] are a few supervised classification methods that have proved to be useful in astrophysics.

2.2 Unsupervised classification

Unsupervised classification techniques search for similarities in the data, usually grouping them into clusters. No training set is needed and no classes are specified before hand, so it is useful for discovering new groups or outliers. After the groups are found, it manual approach is required to interpret the groups or clusters in terms of astrophysical knowledge.

The simplest unsupervised classification method is *K-means* which aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean according to some predefined metrics (commonly euclidean).

2.3 Common techniques

Several of the classification methods do not allow continuous, numeric data. It is mandatory to *discretize* the datasets prior to their use. Most simple techniques just divide the value ranges into a fixed number of bins, or into bins with the same number of elements in them. There also exist supervised methods that take into account the class value in order to make the partitions, being [2] one of the most used. [11] provides a good introduction to several discretization methods.

Irrelevant, redundant, useless or contradictory attributes can degrade the efficiency of the classification system. *Feature selection* methods (FS), while not vital, can provide a boost in the results. We must use heuristics when dealing with a large number of attributes as

finding the best attribute subset is an NP-complete problem (it cannot be solved in polynomial time in any known way). If we have decided before hand the classification algorithm to use, we can use the classifier performance to select the optimal feature subset (i.e., that which produces the lowest misclassification rate). If no committment is to be made prior to feature selection, a so-called filter Feature Selections has to be performed where attributes or attributes sets are ranked according to the relevance, measure with, for example, the entropy, correlation with the class, or mutual information. See [3] for a wide overview.

When the supervised classification is to operate with a large number of classes (typically 10 or more), it is often useful to carry out the classification in a multi-stage scheme whereby the full classification takes place in a number of individual stages with a small number of categories each. The categories in the first steps of the classifier comprise several of the original classes. This approach allows for smaller, more specialized classifiers, and usually improves the results by allowing the use of different classification algorithms and attributes sets in each node. Classification hierarchies have been used intensively in text and web domains (e.g., [1]) where we can have hundreds of class labels.

3 Results

Corot[3] stands for "Convection, Rotation and planetary Transits". "Convection and rotation" refers to the capability of Corot to probe stellar interiors studying the acoustic waves that ripple across the surface of stars (asteroseismology). "Transit" refers to the technique by which the presence of a planet orbiting a star can be inferred from the dimming starlight, caused when the planet passes in front of it. To accomplish these two scientific objectives Corot will monitor about 120 000 stars.

The SVO provides both traditional web-based and VO access to Corot data (see Fig. 1). In collaboration with the Katholieke Universiteit Leuven we also provide information about stellar variability. The main algorithms used to obtained are:

- Bayesian networks and Gaussian classifiers.
- Supervised discretization.
- Supervised feature selection [4].

For a deeper insight into the classification system see [8].

We have used data mining techniques to analyze the correspondence between the classical stellar variability types and the clusters found in the distribution of light curve parameters and colour indices of stars in the Corot exoplanet sample. Our findings have been incorporated in the classification scheme to improve the supervised classification used in the Corot catalogue production, and further improvements will be added once the existence of new classes or subtypes are confirmed by complementary spectroscopic observations [9].

[3] www.esa.int/esaMI/COROT/index.html

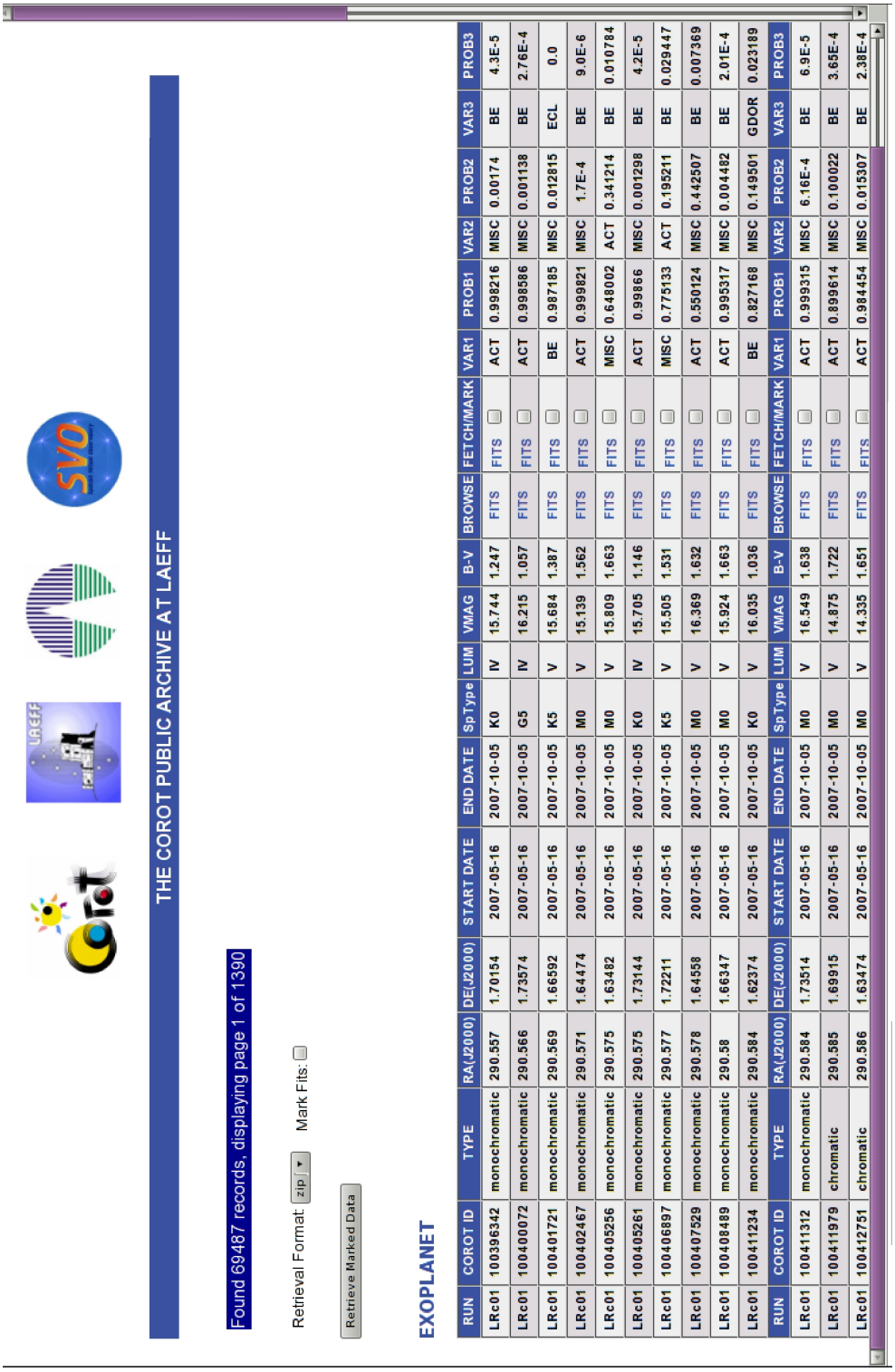

Figure 1: The Corot web service provided by the SVO. It provides access to light curves and several high-level data, such as a target most probable variability classes.

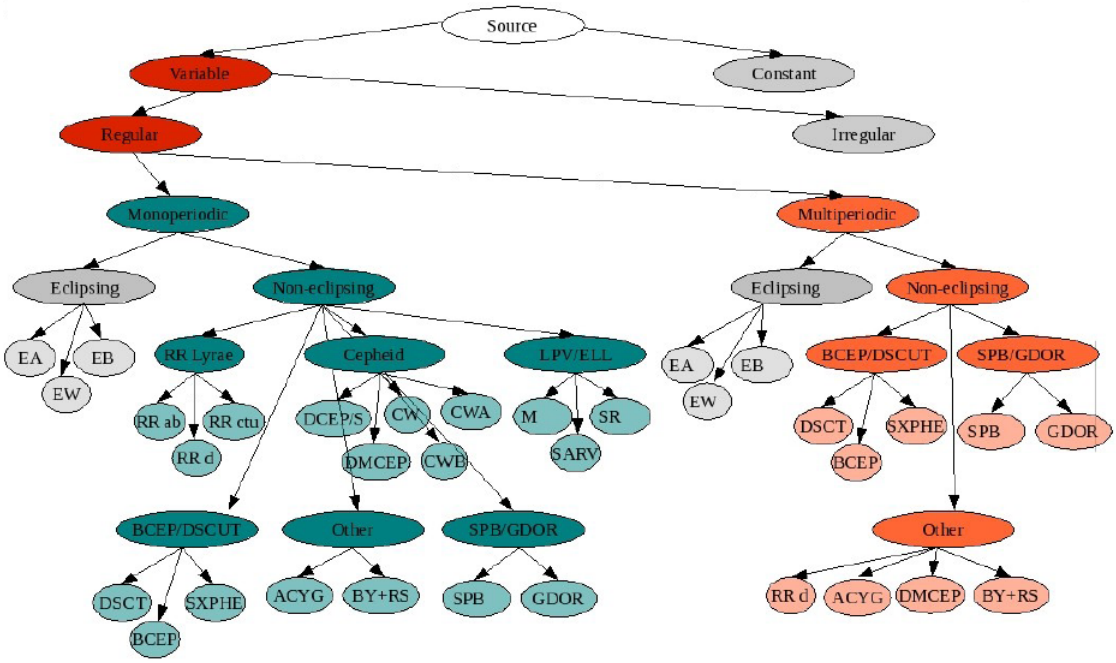

Figure 2: Current Gaia hierarchy for the classification system.

GAIA[4] is an ESA mission which aims to create the largest and most precise three dimensional chart of our Galaxy by providing unprecedented positional and radial velocity measurements for about one billion stars in our Galaxy and throughout the Local Group. Its launch is foreseen for late 2012.

The SVO collaborates with the Katholieke Universiteit Leuven and the INTEGRAL Science Data Centre (ISDC) into the development of a classification system for the data to come. The high number of stars (about 10^8 are expected to be variable stars at the Gaia photometric precision) forces very fast classification algorithms. In particular, we have developed a new algorithm for finding the best hierarchy automatically (which can be useful if new classes are detected), testing the use of error bars in training/test data and handling missing attributes. Figure 2 shows the current hierarchy for Gaia.

References

[1] Chuang, S., & Chien, L. 2005, ACM Journal

[2] Fayyad, U., & Irani, K. 1993, 13th International Joint Conference on Artificial Intelligence, p. 1022

[3] Guyon, I., & Elisseeff, A. 2003, Jorunal of Machine learning Research 3, 1157

[4] sci.esa.int/science-e/www/area/index.cfm?fareaid=26

[4] Hall, M. 1998, Ph. D. thesis, Department of Computer Science, University of Waikato

[5] Li-Li, L. Yan-Xia Z., Yong-Heng, Z., & Da-Wei, Y. 2007, Chin. J. Astron. Astrophys., 7, 448

[6] López, M., Bielza, C., & Sarro, L. M. 2006, ADASS XV, 161,164

[7] Raquel, A., Cecilia, R., & Chris, D. 2006, Proceedings of the 5th International Conference of Machine Learning Applications

[8] Sarro, L. M., Debosscher J., López, M., & Aerts, C. 2009, A&A, 494, 739

[9] Sarro, L. M., Debosscher, J., Aerts, C., & López, M. 2009, A&A, 506, 535

[10] Vasconcellos, E., de Carvalho, R., Gal, R., et al. 2011, AJ, 141, 189

[11] Yang, Y., & Webb, G. 2002, Pacific Rim Knowledge Acquisition Workshop, 159