Star Classification Under Data Variability: An Emerging Challenge in Astroinformatics

Ricardo Vilalta[1(✉)], Kinjal Dhar Gupta[1], and Ashish Mahabal[2]

[1] Department of Computer Science, University of Houston, Houston, TX 77204, USA
{vilalta,kinjal13}@cs.uh.edu
[2] Department of Astronomy,
California Institute of Technology, Pasadena, CA 91125, USA
aam@astro.caltech.edu

Abstract. Astroinformatics is an interdisciplinary field of science that applies modern computational tools to the solution of astronomical problems. One relevant subarea is the use of machine learning for analysis of large astronomical repositories and surveys. In this paper we describe a case study based on the classification of variable Cepheid stars using domain adaptation techniques; our study highlights some of the emerging challenges posed by astroinformatics.

Keywords: Astroinformatics · Domain adaptation · Variable star classification

1 Introduction

The recent emergence of a new field of study named *astroinformatics*, comes as a response to the rapid growth of data volumes corresponding to a variety of astronomical surveys. Data repositories have gone from gigabytes into terabytes, and we expect those repositories to reach the petabytes in the coming years. This massive amount of data stands in need of advanced computational solutions. The general aim in astroinformatics is the application of computational tools to the solution of astronomical problems; key issues involve not only an efficient management of data resources, but also the design of new computational tools that efficiently capture the nature of astronomical phenomena.

Recent work reports on successful applications of machine learning for analysis of large astronomical repositories and surveys [2]. Machine learning is already an indispensable resource to modern astronomy, serving as an instrumental aid during the process of data integration and fusion, pattern discovery, classification, regression and clustering of astronomical objects. And we expect machine learning to produce high-impact breakthroughs when large (petabyte) datasets become available. As an illustration, LSST (Large Synoptic Survey Telescope), will survey the sky to unprecedented depth and accuracy at an impressive temporal cadence [3]; it will generate an expected 30 terabytes of data obtained each night to provide a complete new view of the deep optical universe in the realm of time domain

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 241–244, 2015.
DOI: 10.1007/978-3-319-23461-8_22

astronomy. Other projects, such as CRTS (Catalina Realtime Transient Survey), have already begun to yield impressive scientific results in time-domain astronomy. And projects such as GAIA[1] and DES (Dark Energy Survey) promise to illuminate unprecedented amounts of the time-varying Universe.

2 Cepheid Variable Star Classification

One important challenge in the analysis of astronomical data is that as we move from one survey to another, the nature of the light-curves changes drastically. As an example, some surveys contain rich sources of data in terms of temporal coverage, but the depth is shallow. Other surveys capture objects at extreme depths but for a short time only. And even if we remain within the same survey, analyzing objects that belong to different regions of the sky can bring substantial differences in measurements. All these factors lead to different aspects of data variability.

We now describe a case study where we address the analysis of a rich variety of large surveys under data variability (previous reports can be found in [5,6]). The problem we address is characterized by an original source surveys where class labels for astronomical objects abound, and by a target survey with few class labels, and where feature descriptions may differ significantly (i.e., where marginal probabilities may differ). The problem is also known as *domain adaptation*, or *concept shift*, in machine learning [1,4]. A solution to this common problem in astronomy carries great value when dealing with large datasets, as it obviates the compilation of class labels for the new target set.

Our study is confined to the context of Cepheid variable star classification [5,6], where the goal is to classify Cepheids according to their pulsation modes (we focus on the two most abundant classes, which pulsate in the fundamental and first-overtone modes). Such classification can in fact be attained for nearby galaxies with high accuracy (e.g., Large Magellanic Cloud) under the assumption of class-label availability. The high cost of manually labeling variable stars, however, suggests a different mode of operation where a predictive model obtained on a data set from a source galaxy T_{tr}, is later used on a test set from a target galaxy T_{te}. Such scenario is not straightforwardly attained, as shown in Fig. 1 (left), where the distribution of Cepheids in the Large Magellanic Cloud LMC galaxy (source domain, top sample), deviates significantly from that of M33 galaxy (target domain, bottom sample). In this example, we employ two features only: apparent magnitude in the y-axis, and log period in the x-axis, but our solution is general and allows for a multi-variate representation. Both the offset along apparent magnitude[2], and the significant degree of sample bias, are mostly due to the fact that M33 is $\sim 16\times$ farther than the LMC. Our assumption is then

[1] Satellite mission launched in 2013 by the European Space Agency to determine the position and velocity of a billion stars, creating the largest and most precise 3D map of the Milky Way.

[2] Apparent magnitude m, is defined as $m = -2.5 \times \log_{10} \frac{L}{d^2}$, where d is the distance from Earth to the star measured in parsecs, and L is the star luminosity. Hence, smaller numbers correspond to brighter magnitudes (higher fluxes).

that the difference in the joint input-output distribution between the target and source surveys is mainly due to a systematic shift of sample points.

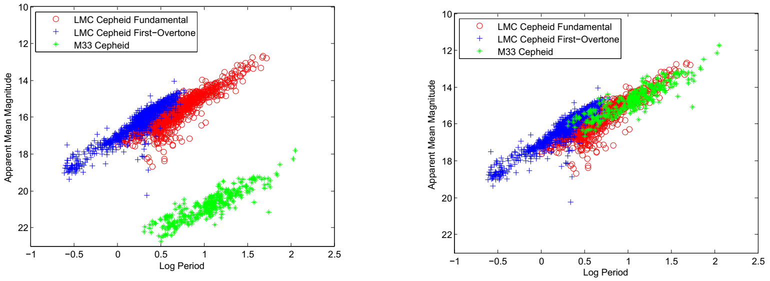

Fig. 1. Left. The distribution of Cepheids along the Large Magellanic Cloud LMC (top sample), deviates significantly from M33 (bottom sample). Right. M33 is aligned with LMC by shifting along mean magnitude.

Our proposed solution shows evidence of the usefulness of domain adaptation in star classification [6]. The main idea consists of shifting T_{te} using maximum likelihood. As an example, if we assume the marginal distribution from which the training is drawn, $P_{\text{tr}}(\mathbf{x})$, is a mixture of Gaussians, we can then estimate parameters directly from our sample T_{tr}, since we know all class labels (i.e., we know which vector belongs to each component or Gaussian). This enables us to have a complete characterization of the marginal distribution: $P_{\text{tr}}(\mathbf{x}) = \sum_{i=1}^{c} \phi_i \;\, g_i(\mathbf{x}|\mu_i, \Sigma_i)$, $g_i(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\}$, where ϕ_i, μ_i, and Σ_i are the mixture coefficient (i.e., prior probability), mean and covariance matrix of the ith component respectively, n is the number of features, and c is the number of components. We can then define a new testing set $T'_{\text{te}} = \{\mathbf{x}'\}$, where $\mathbf{x}' = (x_1 + \delta_1, x_2 + \delta_2 + ... + x_n + \delta_n)$, since we know a shift has occurred along our input features. Our approach is then to find the set of shifts $\Delta = \{\delta_i\}$ that maximizes the log likelihood of T'_{te} with respect to distribution $P_{\text{tr}}(\mathbf{x})$: $\mathcal{L}(\Delta|T'_{\text{te}}) = \log \prod_{k=1}^{q} P_{\text{tr}}(\mathbf{x}^k) = \sum_{k=1}^{q} \log P_{\text{tr}}(\mathbf{x}^k)$.

To solve this optimization problem, we used an iterative gradient ascent approach; we search the space of values in Δ for which the log-likelihood function reaches a maximum value. Fig. 2 shows our results; we used Cepheid variables from Large Magellanic Cloud (LMC) as the source domain, and M33 as the target domain. There is a significant increase in accuracy with the data alignment step, which serves as evidence to support our approach.

3 Conclusions and Remarks

The variability of surveys in terms of depth and temporal coverage in astronomy calls for specialized techniques able to learn, adapt, and transfer predictive models from source light-curve surveys to target light-curve surveys. In this paper we

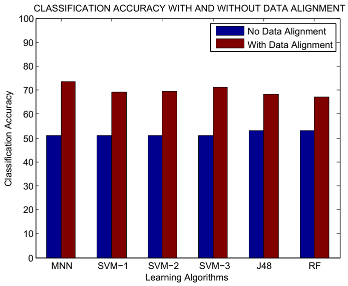

Fig. 2. Plot bars showing classification accuracy when a predictive model built using data from LMC (Large Magellanic Cloud) galaxy is tested on data from M33 galaxy. Blue bars show results when no data alignment is used; red bars show results using data alignment.

show a methodology along this direction that accounts for a data misalignment caused by a systematic data shift.

To end, we point to the importance of exploiting contextual information when modeling astronomical phenomena. This is because the surroundings of a variable source are essential to determine the nature of the object under study. For example, a supernova is easiest to distinguish from other variable and normal objects because it exhibits one brightening episode and then it fades away over weeks. However, if the context reveals the presence of a galaxy nearby, the supernova interpretation becomes much more plausible. A radio source in close proximity to a transient, in contrast, suggests a Blazar classification and is evidence against a supernova. Such contextual information is key to attain accurate predictions, and will become increasingly accessible with the advent of extremely large astronomical surveys.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Wortman, J.: A Theory of Learning from Different Domains. Machine Learning, Special Issue on Learning From Multiple Sources **79**, 151–175 (2010)
2. Ivezic, Z., Connolly, A.J., VanderPlas, J.T., Gray, A.: Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, Princeton Series in Modern Observational Astronomy. Princeton University Press (2014)
3. LSST Science Book, version 2.0, 245 authors. http://www.lsst.org/lsst/scibook
4. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. D.: Dataset Shift in Machine Learning. MIT Press (2009)
5. Vilalta, R., Dhar Gupta, K., Macri, L.: A Machine Learning Approach to Cepheid Variable Star Classification using Data Alignment and Maximum Likelihood. Astronomy and Computing **2**, 46–53 (2013). Elsevier
6. Vilalta, R., Dhar Gupta, K., Macri, L.: Domain adaptation under data misalignment: an application to cepheid variable star classification. In: The 22nd International Conference on Pattern Recognition (ICPR 2014), Stockholm, Sweden (2014)