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...
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\textsuperscript{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\textsuperscript{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

\textsuperscript{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.

\textsuperscript{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.

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, $P_{tr}(x)$, 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:

$$P_{tr}(x) = \sum_{i=1}^{c} \phi_i g_i(x|\mu_i, \Sigma_i),$$

where $\phi_i$, $\mu_i$, and $\Sigma_i$ are the mixture coefficient (i.e., prior probability), mean and covariance matrix of the $i$th component respectively, $n$ is the number of features, and $c$ is the number of components. We can then define a new testing set $T'_{te} = \{x'\}$, where $x' = (x_1 + \delta_1, x_2 + \delta_2 + ... + x_n + \delta_n)$, 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 $P_{tr}(x)$:

$$\mathcal{L}(\Delta|T'_{te}) = \log \prod_{k=1}^{q} P_{tr}(x^{k}) = \sum_{k=1}^{q} \log P_{tr}(x^{k}).$$

To solve this optimization problem, we used an iterative gradient ascent approach; we search the space of values in $\Delta$ 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
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

The Evolution of Social Relationships and Strategies Across the Lifespan

Yuxiao Dong¹, Nitesh V. Chawla¹(✉), Jie Tang², Yang Yang², and Yang Yang¹

¹ Interdisciplinary Center for Network Science and Applications,
Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, USA
{ydong1,nchawla,yyang1}@nd.edu
² Department of Computer Science and Technology,
Tsinghua University, Beijing, China
jietang@tsinghua.edu.cn, yyang.thu@gmail.com

Abstract. In this work, we unveil the evolution of social relationships across the lifespan. This evolution reflects the dynamic social strategies that people use to fulfill their social needs. For this work we utilize a large mobile network complete with user demographic information. We find that while younger individuals are active in broadening their social relationships, seniors tend to keep small but closed social circles. We further demonstrate that opposite-gender interactions between two young individuals are much more frequent than those between young same-gender people, while the situation is reversed after around 35 years old. We also discover that while same-gender triadic social relationships are persistently maintained over a lifetime, the opposite-gender triadic circles are unstable upon entering into middle-age. Finally we demonstrate a greater than 80% potential predictability for inferring users’ gender and a 73% predictability for age from mobile communication behaviors.

Our study [1] is based on a real-world large mobile network of more than 7 million users and over 1 billion communication records, including phone calls and text messages (CALL and SMS). Previous work shows that human social strategies used by people to meet their social needs indicate complex, dynamic, and crucial social theories [2]. This work unveils the significant social strategies and social relationship evolution across one’s lifespan in human communication. Specifically, we investigate the interplay of demographic characteristics and three types of social relationships, including social ego, social tie, and social triad.

The social strategies that people use to build their ego social networks are observed from Figure 1. The X-axis represents central users’ age from 18 to 80 years old and the Y-axis represents the demographic distribution of users’ friends, in which positive numbers denote female friends’ age and negative numbers denote male friends’. The spectrum color, which extends from dark blue

This work was originally published in the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’14) [1]. This extended abstract has been largely extracted from the publication.
(a) Demog. dist. of Female’s friends (b) Demog. dist. of Male’s friends

Fig. 1. Friends’ demographic distribution in ego social networks. X-axis: (a) female age; (b) male age. Y-axis: age of friends (positive: female friends, negative: male friends). The spectrum color represents the friends’ demographic distribution.

(low) to red (high), represents the probability of one’s friends belonging to the corresponding age (Y-axis) and gender (positive or negative). First, the highlighted diagonal lines indicate that people tend to communicate with others of both similar age and gender, i.e., age homophily and gender homophily. Furthermore, we can see that young and middle-age people put increasing focus on the same generation and decreasing focus on the older generation, while the seniors devote more attention on the younger generation even along with the sacrifice of age homophily. Third, we observe that young people are active in broadening social circles (high degree and low clustering coefficient centralities), while seniors tend to keep small but stable connections (low degree and high clustering coefficient centralities).

We further study the social strategies by which people maintain their social tie relationships. The heat maps in Figure 2 visualize the communication frequencies—the number of calls per month between two people with different demographic profiles. Four sub-figures detail the average numbers of calls between two individuals, two males, two females, and one male and one female, respectively. We can see that the interactions between two young males are more frequent than those between two young females (Cf. Figures 2(b) and 2(c)), and moreover, opposite-gender interactions between one young female and male are much more frequent than those between same-gender individuals (Cf. Figure 2(d)). However, reversely, same-gender interactions between two middle-age individuals are more frequent than those between opposite-gender
Fig. 2. **Strength of social tie.** XY-axis: age of users with specific gender. The spectrum color represents the number of calls per month. (a), (b), and (c) are symmetric.

More interestingly, we highlight the social strategies on triadic relationships unveiled from Figure 3, wherein the X-axis and Y-axis denote the minimal and maximal age of three users within a closed social triad. Sub-figures 3(a) and 3(d) show the distributions of same-gender triads: ‘FFF’ (Female-Female-Female) and ‘MMM’ (Male-Male-Male), and sub-figures 3(b) and 3(c) present distributions for users’ age in opposite-gender triads: ‘FFM’ and ‘FMM’. From heat-map visualization, we observe that people expand both the same-gender and opposite-gender triadic relationships during the dating active period. However, people’s attention to opposite-gender circles quickly disappears after entering into middle-age (Cf. Figures 3(b) and 3(c)) and the same-gender triadic relationships are persistent over a lifetime (Cf. Figures 3(a) and 3(d)). To the best of our knowledge, we are the first to discover the instability of opposite-gender triadic relationships and the persistence of same-gender triadic relationships over a lifetime in a large
population, which demonstrates the evolution of social strategies that are used by people to meet their social needs in different life stages.

Based on these discovered social strategies, we further study to what extent users’ demographic information can be inferred from mobile communication behaviors. The objective is to infer users’ gender and age simultaneously by leveraging their interrelations. We present the WhoAmI framework—a Multiple Dependent-Variable Factor Graph model, whereby the social interrelations between users with different demographic profiles can be modeled. On both CALL and SMS networks, the WhoAmI method can achieve an accuracy of 80% for predicting users’ gender and 73% for users’ age according to their daily mobile communication patterns, significantly outperforming several alternative data mining methods.

**Fig. 3.** Demographic distribution in social triadic relationships. X-axis: minimum age of three users in a triad. Y-axis: maximum age of three users. The spectrum color represents the distributions.

**Acknowledgments.** Nitesh V. Chawla, Yuxiao Dong, and Yang Yang are supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405. Jie Tang
and Yang Yang are supported by the National High-Tech R&D Program of China (No. 2014AA015103), Natural Science Foundation of China (No. 61222212), and National Basic Research Program of China (2014CB340506).

References