Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314013775

Adapting Predictive Models for Cepheid Variable Star Classification Using Linear Regression and Maximum Likelihood

Kinjal Dhar Gupta[1], Ricardo Vilalta[1], Vicken Asadourian[2] and Lucas Macri[3]

[1]Dept. of Computer Science, University of Houston.

[2]Dept. of Mathematics, University of Houston.
4800 Calhoun Road , Houston TX-70004, USA.
email: kinjal13@cs.uh.edu, vilalta@cs.uh.edu, vmasadourian@uh.edu

[3]Dept. of Physics and Astronomy, Texas A&M University.
4242 TAMU , College Station, TX 77843-4242, USA.
email: lmacri@tamu.edu

Abstract. We describe an approach to automate the classification of Cepheid variable stars into two subtypes according to their pulsation mode. Automating such classification is relevant to obtain a precise determination of distances to nearby galaxies, which in addition helps reduce the uncertainty in the current expansion of the universe. One main difficulty lies in the compatibility of models trained using different galaxy datasets; a model trained using a training dataset may be ineffectual on a testing set. A solution to such difficulty is to adapt predictive models across domains; this is necessary when the training and testing sets do not follow the same distribution. The gist of our methodology is to train a predictive model on a nearby galaxy (e.g., Large Magellanic Cloud), followed by a model-adaptation step to make the model operable on other nearby galaxies. We follow a parametric approach to density estimation by modeling the training data (anchor galaxy) using a mixture of linear models. We then use maximum likelihood to compute the right amount of variable displacement, until the testing data closely overlaps the training data. At that point, the model can be directly used in the testing data (target galaxy).

Keywords. (stars: variables:) Cepheids, (galaxies:) Magellanic Clouds, methods: statistical, infrared: stars, methods: data analysis.

1. Introduction

Traditional machine learning algorithms assume both training and testing data originate from the same distribution. This comes unwarranted in real-world applications. One approach to handle the discrepancy between source (training) and target (test) domains is called *domain adaptation*, where class-conditional distributions remain equal, though class prior distributions differ Ben-David *et al.* (2006), Storkey (2009), Ben-David *et al.* (2010). Our domain adaptation method learns a model on a source domain without using any information from a target domain; we assume equal class conditional probabilities, but class priors differ by a certain shift across one or more features as explained by Vilalta *et al.* (2013). Different from previous work, we assume a bi-variate Linear Mixture Model with Gaussian noise, and use Maximum Likelihood to find the shift between source and target distributions. The idea is to align the two datasets so that a model learnt on the source domain can be effectively used on the target domain.

We classify a particular type of variable stars named Cepheids into two pulsation modes. We use two features: Apparent Mean Magnitude and Logarithm of Period, and

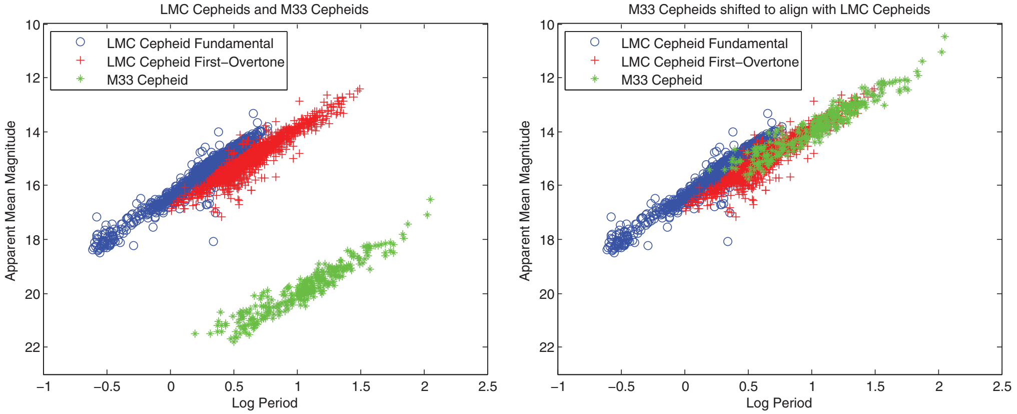

Figure 1. Left. The distribution of Cepheids along the Large Magellanic Cloud LMC (top sample), deviates significantly from M33 (bottom sample). Right. M33 is aligned with LMC by shifting along Mean Magnitude.

two classes: Fundamental and First-Overtone. In our experiments, we take the Large Magellanic Cloud (LMC) (Fig. 1a) as our source domain and M33 (Fig. 1b) as our target domain. All data have been obtained from the Optical Gravitational Lensing Experiment (OGLE) III in the infrared band.

2. Mathematical Formulation

We generate a mixture of linear regression models (as described in Faria & Soromenho (2010)). The log likelihood equation for LMC data for the two features X (Log Period) and Y (Apparent Mean Magnitude) can be written as follows:

$$LogL(\theta|y_1, y_2, ...y_n, x_1, x_2, ...x_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{2} \pi_j \phi_j(y_i|x_i)\right) \tag{2.1}$$

where $y_1, y_2, ...y_n$ are observations of Y, $x_1, x_2, ...x_n$ are observations of X, and each component $\phi_j(y_i|x_i)$ corresponds to a Gaussian distribution $\mathcal{N}(x_i^T\beta_j, \sigma_j)$ with coefficient π_j. Parameter estimates are captured in θ. The equation can be expanded as follows:

$$LogL(\theta|y_1, y_2, ...y_n, x_1, x_2, ...x_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{2} \pi_j \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(y_i - x_i^T\beta_j)^2}{2\sigma_j^2}}\right) \tag{2.2}$$

where $x_i = [x_{i1}\ x_{i2}]^T$ and $\beta_j = [\beta_{j1}\ \beta_{j2}]^T$. Here x_{i1} and x_{i2} indicate the value of Log Period and the bias variable for the ith observation respectively (x_{i2}=1 for all observations). β_{j1} and β_{j2} are the corresponding regressor variables for the jth component. To align M33 data with LMC data, we assume a shift δ across Mean Magnitude only. The log likelihood equation for M33 data can be written as follows:

$$LogL(\theta, \delta|y_1, y_2, ...y_n, x_1, x_2, ...x_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{2} \pi_j \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(y_i + \delta - x_i^T\beta_j)^2}{2\sigma_j^2}}\right) \tag{2.3}$$

To find the maximum value of δ we differentiate Eq. (2.3) with respect to δ and equate to zero. After some algebraic manipulation we derive the following equation:

$$\left(\frac{c_{i12}\alpha_{i1}e^{-\gamma_{i1}(\delta)}+c_{i22}\alpha_{i2}e^{-\gamma_{i2}(\delta)}}{\alpha_{i1}e^{-\gamma_{i1}(\delta)}+\alpha_{i2}e^{-\gamma_{i2}(\delta)}}\right)+\left(\frac{c_{i13}\alpha_{i1}e^{-\gamma_{i1}(\delta)}+c_{i23}\alpha_{i2}e^{-\gamma_{i2}(\delta)}}{\alpha_{i1}e^{-\gamma_{i1}(\delta)}+\alpha_{i2}e^{-\gamma_{i2}(\delta)}}\right)=0 \qquad (2.4)$$

where

$$c_{ij1}=\frac{(y_i+\delta-x_i^T\beta_j)^2}{2\sigma_j^2},\qquad c_{ij2}=\frac{(y_i+\delta-x_i^T\beta_j)}{\sigma_j^2},\qquad c_{ij3}=\frac{1}{2\sigma_j^2},$$

$$\alpha_{ij}=\pi_j\frac{1}{\sigma_j\sqrt{2\pi}}e^{-c_{ij1}},\qquad \gamma_{ij}(\delta)=c_{ij2}\delta+c_{ij3}\delta^2$$

3. Experiment Results

Fitting a Mixture of Linear Models. We use Eq. (2.2) to fit a mixture of linear models to the source data. The two components refer to the two classes: Fundamental and First-Overtone. We use the EM algorithm (Faria & Soromenho 2010) to do parameter estimation. The EM algorithm stops when the change in Q-value in the E-step falls below 10^{-10}. Initial values and final parameter estimates are shown in Table 1.

Aligning M33 Data with LMC Data. We use the parameter values in Table 1 to find the shift δ between LMC and M33. Solving Eq. (2.6) yields a value of $\delta = -6.06$. After shifting M33 data by δ we obtain the results shown in Table 2. An asterisk denotes a significant difference. Experimental results show how classification accuracy increases significantly after the source and target datasets are aligned.

Table 1. Parameter Estimates

Parameter	Initial Value	Final Value
β_1	$\begin{bmatrix}-3.259\\16.407\end{bmatrix}$	$\begin{bmatrix}-3.209\\16.358\end{bmatrix}$
β_2	$\begin{bmatrix}-2.969\\16.904\end{bmatrix}$	$\begin{bmatrix}-2.929\\16.889\end{bmatrix}$
σ_1	0.1	0.164
σ_2	0.1	0.214
π_1	0.404	0.384
π_2	0.596	0.616

Table 2. Classification Accuracies

Learning Algorithm	Without Shift	With Shift
Neural Networks	92.70(1.19)	97.90(0.89)*
Support Vector Machine with Polynomial Kernel 1	92.70(1.19)	96.27(0.86)*
Support Vector Machine with Polynomial Kernel 2	92.70(1.19)	96.10(0.80)*
Support Vector Machine with Polynomial Kernel 3	92.70(1.19)	96.53(0.88)*
J48 Decision Tree	92.90(1.24)	98.07(0.83)*
Random Forest	92.90(1.24)	98.03(0.69)*

References

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. 2006, *Neural Information Processing Systems*, 19, 137-144

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., & Wortman, J. 2010, *Machine Learning*, 79, 151-175

Faria, S. & Soromenho, G. 2010, *Journal of Statistical Computation and Simulation*, Vol. 80, No. 2, 201-225

Storkey, A. 2009, *Dataset Shift in Machine Learning, MIT Press*, 3-28

Vilalta, R., Dhar Gupta, K., & Macri, L. 2013, *Astronomy and Computing*, 2, 46-53

https://doi.org/10.1017/S1743921314013775 Published online by Cambridge University Press