Finite Dirichlet Mixture Modeling for Classification and Detection of new classes of Variable stars

Prince John, Dr. Maria Süveges and Prof. Alessandra R Brazzale

Università degli Studi di Padova

MOTIVATION

The GAIA survey is observing about 1 billion stars, which is roughly one percent of the total stellar content of the Milky Way. Consequently, a huge influx of Variable star data is expected to arrive, which will need to be classified into different classes. There are several classes of variable stars; 23 in our training dataset (Dubath et. al, 2011). The aim of our work is

- **Goal 1 :** To implement a statistical model which can classify the awaited GAIA dataset into these 23 classes.
- **Goal 2 :** To implement a statistical model which can detect new classes of variable stars in the GAIA dataset.

METHODS AND MODELS

TWO STAGE MODEL

In connection with Goal 1, we propose a *Two Stage Dirichlet model* by fitting each of the 23 classes of variable stars to a finite mixture of Dirichlet distributions and then forming an ensemble of all the 23 mixtures, which is represented as follows.

For a D dimensional dataset $x = (x_1, x_2, \ldots, x_D)$ which belongs to the $D-1$ dimensional simplex and $x_D = 1 - \sum_{i=1}^{D-1} x_i$

$$f(x) = \sum_{i=1}^{M} \rho_i \sum_{j=1}^{n_i} \pi_{ij} \frac{1}{B(\alpha_{ij})} \prod_{k=1}^{D} x_k^{\alpha_{ijk}-1}$$

where $\rho_i, i = 1 : M, \rho_i > 0, \sum \rho_i = 1$ are the outer mixture probabilities and $\pi_{ij}, j = 1 : n_i, \pi_{ij} > 0, \sum \pi_{ij} = 1$ are the inner mixture probabilities. Also n_i is the number of inner mixture components for the i^{th} outer mixture and M is the number of variable star classes, M=23 in our case.The Dirichlet parameters for the i^{th} variable class(or outer mixture) and j^{th} inner mixture is $\alpha_{ij} = (\alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijD})$ and $B(\alpha_{ij})$ is the beta function over α_{ij}

NEW CLASS DETECTION

For Goal 2, we propose the use of a semi-supervised classification method (Vatanen et al. 2012) for detecting new classes of variable stars. The model is called the Fixed Background model. The Two Stage Dirichlet model is used as the background model and a finite mixture of Dirichlet distributions is used as the anomaly model. This model can help detect the new classes in the GAIA dataset apart from the 23 classes in the training dataset.

Again for the D dimensional dataset $x = (x_1, x_2, \ldots, x_D)$ which belongs to the $D-1$ simplex and $x_D = 1 - \sum_{i=1}^{D-1} x_i$,

$$f_{FB}(x) = \pi_B p_B(x) + \sum_{q=J-1}^{J+Q} \pi_q Dir(x|\alpha_q)$$

where π_B is the component probability of the fixed background model, $p_B(x)$ and $\pi_{J+1}, \ldots, \pi_{J+q}$ are the component probabilities of the anomaly model, $\sum_{q=J-1}^{J+Q} \pi_q Dir(x|\alpha_q)$. π_q, which is a mixture of Q Dirichlet distributions, with component probabilities π_q

MODIFIED SUPERVISED CLASSIFICATION

Again, in connection with Goal 1, we propose a Modified Supervised Classification technique. We can use the Two Stage model or the New Class detection model for this method.

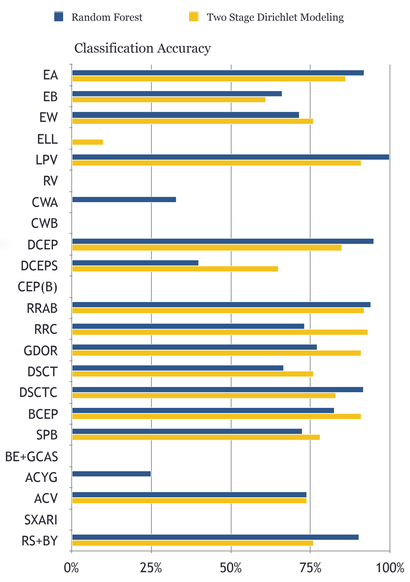
Pre-requisite step : The model (Two Stage/ New Class detection model) is trained by the labeled dataset.

- Step 1 : Unlabeled dataset is fed into the model and classified into different variable classes.
- Step 2 : The labeled dataset and the data which was classified in Step 1 together forms the training data and the model is trained.

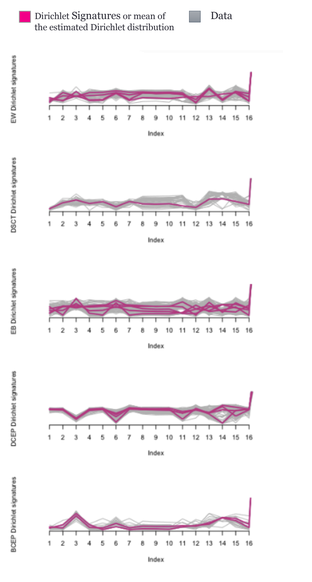
Steps 1 and 2 are repeated whenever a new dataset is fed into the model for classification.

RESULTS

CLASSIFICATION RESULTS



Overall classification accuracy is 69.3% as opposed to the Random forest accuracy of 80%, but for 9 out of 18 classes there is improvement. This trained model can be used to classify the GAIA dataset.

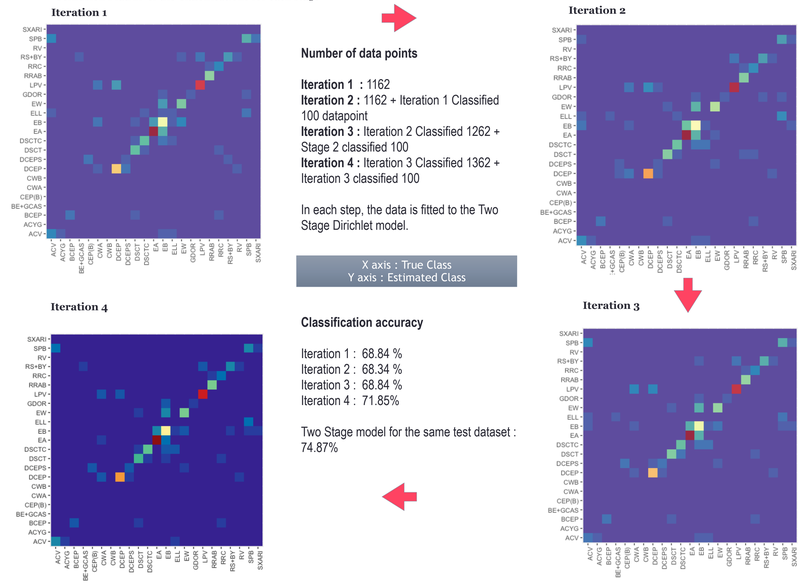


For Dirichlet distribution with parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, the Dirichlet signature is $(\frac{\alpha_1}{\sum_{i=1}^{K} \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{K} \alpha_i}, \ldots, \frac{\alpha_D}{\sum_{i=1}^{K} \alpha_i})$.

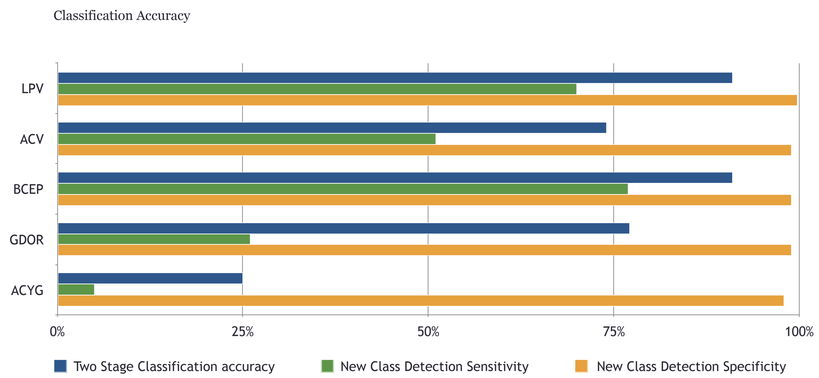
0

MODIFIED SUPERVISED CLASSIFICATION RESULTS

Confusion Matrix of the Classifications for each stage

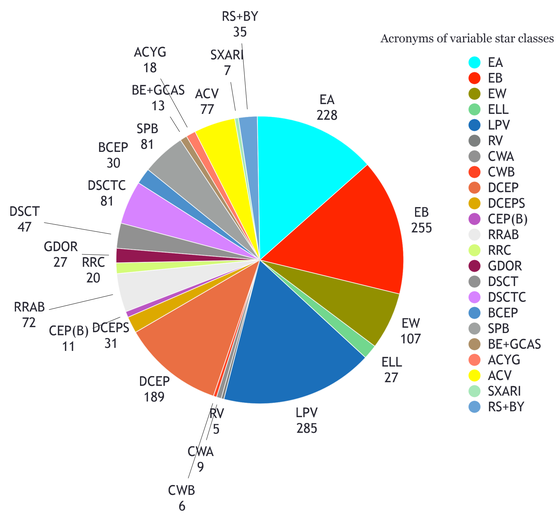


NEW CLASS DETECTION RESULTS



DATASET

The entire dataset contains 1661 data points (Dubath et al, 2011). Out of which 1162 were taken for training and 499 for testing.



FUTURE RESEARCH

- The complete data likelihood of the Two Stage Dirichlet model is a member of the exponential family and also the parameter prior is conjugate to the model. Hence it is a member of the Conjugate Exponential family (Beal, M. and Ghahramani, Z. (2003)). Variational EM algorithm can be used to estimate the parameters.

ESSENTIAL BIBLIOGRAPHY

1) Vatanen, Tommi, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen, and Yoshikazu Nagai. 2012. "Semi-Supervised Detection of Collective Anomalies with an Application in High Energy Particle Physics." In *Neural Networks (IJCNN), the 2012 International Joint*
2) Dubath, P, L Rimoldini, M Süveges, Jonas Blomme, M López, LM Sarro, Joris De Ridder, et al. 2011. "Random Forest Automated Supervised Classification of Hipparcos Periodic Variable Stars." *Monthly Notices of the Royal Astronomical Society* 414 (3). Oxford University Press: 2602–17. Conference on, 1–8. IEEE.