Detection of RR-Lyrae variable stars using traditional machine learning

Jeremias Rodriguez
Universidad Nacional de Rosario
Rosario, Argentina
jeremiaslcc@gmail.com

1 Introduction

Astronomy is undergoing a profound transformation due to the development of superb modern telescopes, which have made it possible for astronomers to conduct major astronomical surveys. Owing to the overwhelming size of the data collected, automated procedures are vital to analyze and extract valuable information from raw telescope observations.

In this project several machine learning and data mining techniques have been used to identify variable stars of type **RR-Lyrae** from infrared telescope observations published by the VVV Survey (2010 – 2015) [1]. Conducted at the Parnal Observatory in Chile, the survey observed approximately 10^9 different variable stars in the Milky Way over a period of 5 years.

Variable stars of RR-Lyrae type are extremely useful for astronomers, as they are a prime tool with which to obtain distances to old stellar populations in the Milky Way [2] [3]. The work here presented builds on the research published by two of my professors, who successfully applied traditional machine learning to this task [4] [5].

This project focused on two main goals. First, improving the predictions obtained by Support Vector Machine (SVM) [7] classifiers. As a second goal, we sought to understand why tree ensemble methods (such as Random Forest [6] (RF)) grossly outperformed other classifiers in this task, especially SVMs [4] [5]. In this extended abstract I will mainly expand of the results of my first goal.

2 The data

For each tile of the VVV survey (see figure 2), a single dataset was laboriously extracted from the raw telescope measurements [4]. On average, each of these tile datasets describes 500000 variable stars using 62 numerical attributes that summarize their light curves, such as its estimated period and pseudo-color.

By cross matching with previous overlapping star catalogs, it was possible to label some variable stars as RR-Lyrae on a handful of tiles. In other tiles, there are very few or no known stars of type RR-Lyrae. The ultimate goal of this research is to train machine learning classifiers on tiles where some stars are labeled as RR-Lyrae, and then apply the trained classifiers to the unlabeled tiles in order to identify new candidate RR-Lyrae stars.

3 Methodology

This project was focused only on four selected tiles, given the sheer size of the data and limited computational resources. The first step was to reproduce the results of my professors' research [4]:

- Using 10-fold grid search cross validation over one of the tiles, the hyperparametrs of Random Forest and Support Vector Machines were optimized (Only linear and RBF kernels were used). The optimized hyperparameters were subsequently used to test the model on the remaining tiles, calculating precision-recall curves as the data is highly imbalanced.

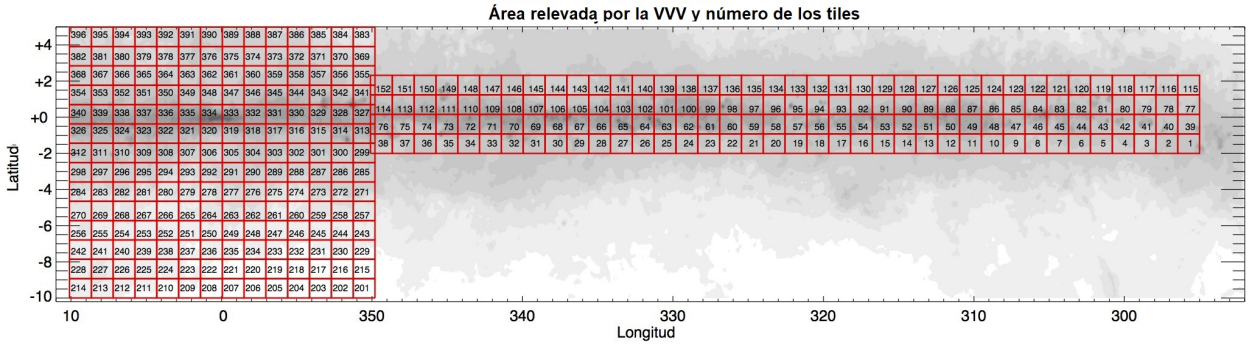

Figure 1: This picture shows the area of the Milky Way that was surveyed in the VVV. The mapped area consists of a total of 348 rectangular sections, or tiles.

- Due to the remarkably large size of the datasets, kernel approximators [8] were used for SVM RBF in order to cope with the huge size of the datasets. This is a novelty compared to previous research, where RBF kernels could not be used in the entirety of the data.
- The results verified that RF outperforms SVM in all the tested combinations of tiles. In the appendix, figure 5 shows a couple of sample precision-recall curves.

Having obtained this baseline, the rest of the project consisted on trying to improve SVM precision-recall curves.

4 Experiments and results

A large number of experiments were conducted, including:

1. **Preprocessing techniques:** Given that the data comes from telescope measurements that were affected by meteorological conditions and measurement errors, some variables are extremely noisy and show a large number of outliers. A wide variety of scaling, transformation and discretization techniques were explored. The results showed a dramatic improvement in SVM after applying binning quantile [8], likely because it smooths out measurement errors and outliers. Trees do this implicitly, and random forest results were not improved by these experiments.
2. **Feature selection and feature extraction:** Some features used for describing variable stars might be too noisy or even not informative at all, since they were empirically selected by astronomers. Experiments were carried out using univariate filters [9], PCA [9] and variable clustering [10]. This further improved the results for SVM, increasing the performance of SVM when selecting 45 out of 63 variables using filters. As part of this analysis, it was interesting to identify which features were less informative, and also to rank the most important features for each classifier [11] [12]. RF seemed to be able to quickly identify the most important features as part of its training process.
3. Other experiments included removing **highly correlated features**, trying different **SVM loss functions**, and reducing the **class imbalance** (both undersampling [13] and synthetic oversampling [14] techniques, as well as class importance [15] approaches). A last experiment attempted to correct the **dataset shift** [16] present between different tiles.

By combining these different techniques, the precision-recall curves of SVM classifiers were able to match RF curves in most tiles tested. The area under the precision-recall curve increased, on average, 22% for SVM classifiers compared to previous works. Two sample precision recall curves can be seen at the appendix, in figure 5.

Extrapolating the final average performance results to the whole set of tiles, the new SVM classifier should be able to identify approximately 30% of the total number of RR-Lyrae stars in the VVV survey with a precision above 90%.

The full code of these experiments is available at `github.com/jere1882/RRL`.

5 Appendix

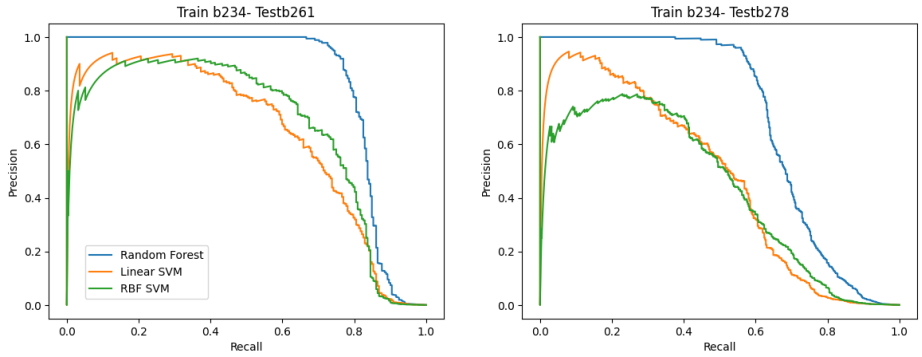

Figure 2: baseline precision-recall curves of RRL classifiers

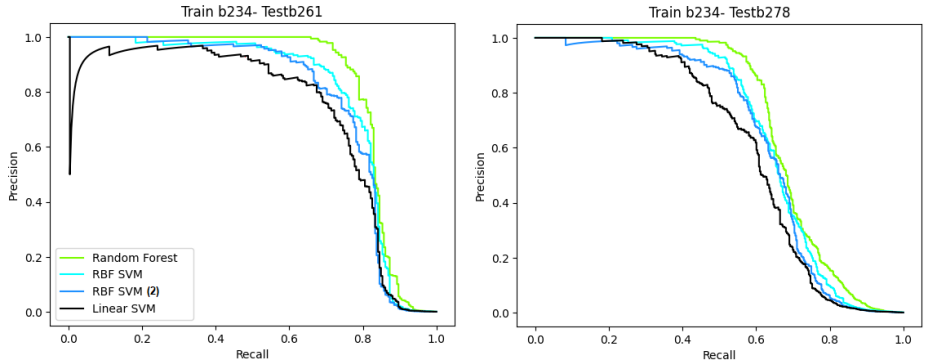

Figure 3: precision-recall curves of RRL classifiers after this project

References

[1] Minniti, D., Lucas, P., Emerson, J., Saito, R., Hempel, M., Pietrukowicz, P., Ahumada, A., Alonso, M., Alonso-Garcia, J., Arias, J., Bandyopadhyay, R., Barb, R., Barbuy, B., Bedin, L., Bica, E., Borissova, J., Bronfman, L., Carraro, G., Catelan, M., and Zoccali, M. (2010). *Vista variables in the via lactea (vvv): The public eso near-ir variability survey of the milky way.* New Astronomy, 15:433.

[2] Shapley, H. (1918). *Studies based on the colors and magnitudes in stellar clusters. vii. the distances, distribution in space, and dimensions of 69 globular clusters.* 48:154-181.

[3] Baade, W. (1946). *A search for the nucleus of our galaxy. Publications of the Astronomical Society of the Pacic*, 58:249.

[4] Cabral, J. B., Ramos, F., Gurovich, S., and Granitto, P. (2020a). Automatic Catalog of RRLyrae from 14 million VVV Light Curves: How far can we go*with traditional machine-learning?* arXiv e-prints, page arXiv:2005.00220.

[5] Elorrieta Lopez, F., Eyheramendy, S., Jordan, A., and Dekany, I. (2016). *A machine learned classiier for rr lyrae in the vvv survey.* Astronomy & Astrophysics, 595.

[6] Breiman, L. (2001). *Random forests.* Machine Learning, 45:5-32.

[7] Cortes, C. and Vapnik, V. (1995). *Support-vector networks.* Mach. Learn., 20(3):273-297.

[8] Williams, C. and Seeger, M. (2001). *Using the nystrom method to speed up kernel machines.* Processing Systems, volume 13, pages 682-688. MIT Press.

[9] Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition.* Morgan Kaufmann Publishers, Waltham, Mass.

[10] ain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[12] Fisher, A., Rudin, C., and Dominici, F. (2018). *Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective*

[13] Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies.* Proceedings of the 2000 International Conference on Artificial Intelligence ICAI.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *Smote: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16:321-357.

[15] Akbani, R., Kwek, S., and Japkowicz, N. (2004). *Applying support vector machines to imbalanced data sets.* volume 3201, pages 39-50.

[16] Quinonero-Candela, J., Sugiyama, M., Lawrence, N., and Schwaighofer, A. (2009). Shift in Machine Learning. Neural information processing series. MIT Press.