A multidimensional genetic programming approach for identifying epsistatic gene interactions

William La Cava*
University of Pennsylvania
Philadelphia, PA, USA
lacava@upenn.edu

Sara Silva
University of Lisbon
University of Coimbra
Universidade Nova de Lisboa
Lisbon, PT

Kourosh Danai
University of Massachusetts Amherst
Amherst, MA, USA

Lee Spector
Hampshire College
Amherst, MA, USA

Leonardo Vanneschi
Universidade Nova de Lisboa
Lisbon, PT

Jason H. Moore
University of Pennsylvania
Philadelphia, PA, USA

ABSTRACT

We propose a novel methodology for binary and multiclass classification that uses genetic programming to construct features for a nearest centroid classifier. The method, coined M4GP, improves upon earlier approaches in this vein (M2GP and M3GP) by simplifying the program encoding, using advanced selection methods, and archiving solutions during the run. In our recent paper, we test this stategy against traditional GP formulations of the classification problem, showing that this framework outperforms boolean and floating point encodings. In comparison to several machine learning techniques, M4GP achieves the best overall ranking on benchmark problems. We then compare our algorithm against state-ofthe-art machine learning approaches to the task of disease classification using simulated genetics datasets with up to 5000 features. The results suggest that our proposed approach performs on par with the best results in literature with less computation time, while producing simpler models.

KEYWORDS

classification, genetics, feature construction

ACM Reference Format:
William La Cava, Sara Silva, Kourosh Danai, Lee Spector, Leonardo Vanneschi, and Jason H. Moore. 2018. A multidimensional genetic programming approach for identifying epsistatic gene interactions. In *GECCO '18 Companion: Genetic and Evolutionary Computation Conference Companion, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3205651.3208217

1 SUMMARY

This paper considers a classification methodology in which genetic programming (GP) is used to contruct feature spaces for distance-based classification. Programs project the original data into a new feature space of potentially different dimensionality, in which a nearest centroid classifier is used to make predictions. The performance improvements observed by earlier versions of this algorithm, M2GP and M3GP, stemmed the incorporation of a distance-based classification strategy into a multi-output GP system. In a recent paper [1], we studied a new method called M4GP, that, although inspired by M2GP and M3GP, significantly improves these two methods. There are 3 methodological contributions of this work. First, M4GP uses a novel (stack-based) program representation that simplifies the construction of multidimensional solutions compared to M2GP and M3GP (which, instead, used a tree-based representation). We demonstrate the effectiveness of this approach in comparison to M2GP and M3GP. Second, M4GP incorporates a multiobjective parent selection and survival technique that allows it to clearly and consistently outperform M2GP and M3GP on a wide set of test problems. To the best of our knowledge, this technique had never been used for multiclass classification before. Third, we introduce an archiving strategy that maintains a set of optimal trade-off solutions based on complexity and accuracy. The final model is selected from this archive using an internal validation set to reduce ovefitting. Thanks to these improvements, M4GP is able to improve the best known GP methods for multi-class classification, and finds results that are competitive with the state-of-the-art methods for 10 benchmark problems.

Most notably, on a set of biomedical data sets with up to 5000 attributes, M4GP and M4GP+EKF (M4GP with feature selection) is shown to perform on par with state-of-the-art methods (XGBoost, deep neural networks and an automated ML approach known as TPOT) while producing smaller models in less time. This is demonstrated in Fig. 1, where 10-fold balanced accuracies of various methods are compared. M4GP does so while using less computation time (Fig. 2), and producing simple feature spaces (Fig 3).

REFERENCES

[1] William La Cava, Sara Silva, Kourosh Danai, Lee Spector, Leonardo Vanneschi, and Jason H. Moore. 2018. Multidimensional genetic programming for multiclass classification. *Swarm and Evolutionary Computation* (2018). https://doi.org/10.1016/j.swevo.2018.03.015

*corresponding author

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3208217

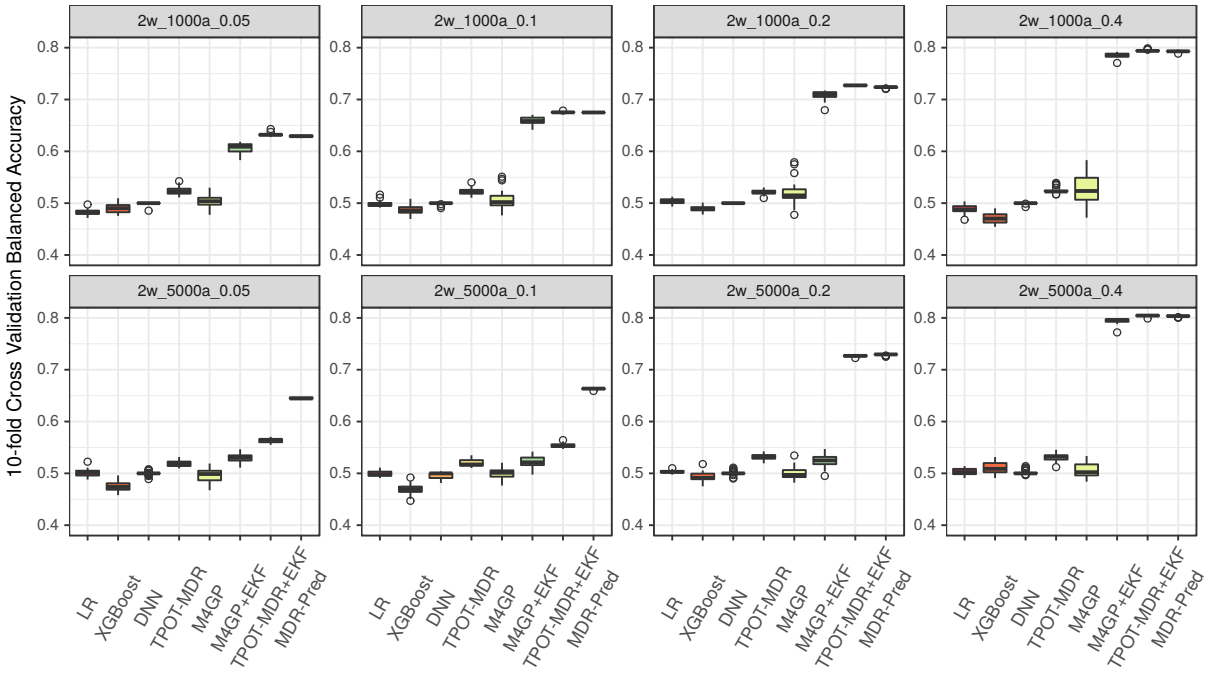

Figure 1: Performance on the biomedical datasets. The subplot titles indicate the dataset; datasets are named by the convention 2w_[# attributes]a_ [signal-to-noise ratio]. Difficulty decreases to the right. MDR-Pred acts as the best possible result.

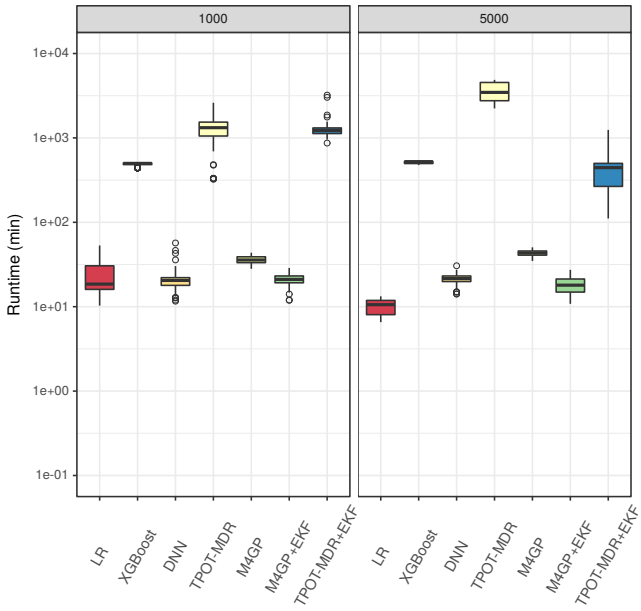

Figure 2: Runtimes on the biomedical datasets. From left to right, the number of attributes in the datasets increase.

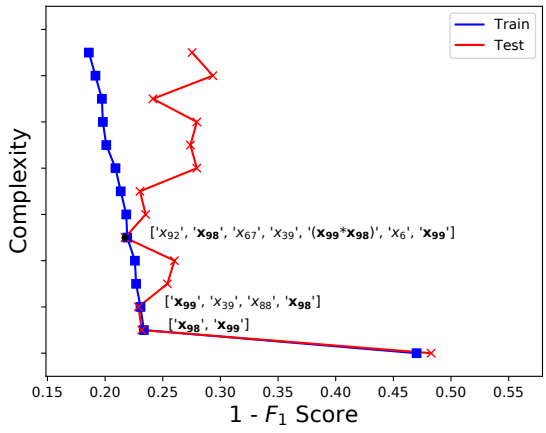

Figure 3: Archive of solutions from M4GP for 2w_100a_0.4her. Blue dots are the training fitness of solutions in the archive, and red is the fitness of those solutions on the holdout test set. The top 3 M4GP solutions are shown with the interacting, predictive genes x_98 and x_99 in bold.