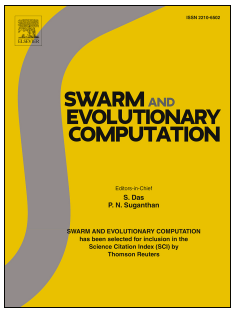
Accepted Manuscript



Multidimensional genetic programming for multiclass classification

William La Cava, Sara Silva, Kourosh Danai, Lee Spector, Leonardo Vanneschi, Jason H. Moore

PII: S2210-6502(17)30913-6

DOI: 10.1016/j.swevo.2018.03.015

Reference: SWEVO 390

To appear in: *Swarm and Evolutionary Computation BASE DATA*

Received Date: 15 November 2017

Revised Date: 26 February 2018

Accepted Date: 30 March 2018

Please cite this article as: W. La Cava, S. Silva, K. Danai, L. Spector, L. Vanneschi, J.H. Moore, Multidimensional genetic programming for multiclass classification, *Swarm and Evolutionary Computation BASE DATA* (2018), doi: 10.1016/j.swevo.2018.03.015.

Multidimensional genetic programming for multiclass classification

William La Cava[a,*], Sara Silva[b,c,d], Kourosh Danai[e], Lee Spector[f], Leonardo Vanneschi[c], Jason H. Moore[a]

[a] *Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[b] *BioISI – Biosystems & Integrative Sciences Institute, Faculty of Sciences, University of Lisbon, Lisbon, Portugal*
[c] *NOVA IMS, Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal*
[d] *CISUC, Department of Informatics Engineering, University of Coimbra, Portugal*
[e] *Dept. of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, USA*
[f] *School of Cognitive Science, Hampshire College, Amherst, USA*

Abstract

We describe a new multiclass classification method that learns multidimensional feature transformations using genetic programming. This method optimizes models by first performing a transformation of the feature space into a new space of potentially different dimensionality, and then performing classification using a distance function in the transformed space. We analyze a novel program representation for using genetic programming to represent multidimensional features and compare it to other approaches. Similarly, we analyze the use of a distance metric for classification in comparison to simpler techniques more commonly used when applying genetic programming to multiclass classification. Finally, we compare this method to several state-of-the-art classification techniques across a broad set of problems and show that this technique achieves competitive test accuracies while also producing concise models. We also quantify the scalability of the method on problems of varying dimensionality, sample size, and difficulty. The results suggest the proposed method scales well to large feature spaces.

Keywords: genetic programming, multiclass classification, feature extraction, feature selection, feature synthesis, dimensionality reduction

1. Introduction

Feature selection and feature construction play fundamental roles in the application of machine learning (ML) to classification. Feature selection makes it possible, for example, to reduce high-dimensional datasets to a manageable size, and to refine experimental designs through measurement selection in some domains. The ML community has become increasingly aware of the need for automated and flexible feature engineering methods to complement the large set of classification methodologies that are now widely available in open-source packages such as Weka and Scikit-Learn [10, 30]. Typical classification pipelines treat feature selection and feature construction as pre-processing steps, in which the attributes in the dataset are selected according to some heuristic [9] and then projected into more complex feature spaces using e.g. kernel functions [29]. In both cases the feature pre-processing is often conducted in a trial-and-error way rather than being automated or intrinsic to the learning method. The use of non-linear feature expansions can also lead to classifiers that are black-box, making it difficult for researchers to gain insight into the modelled process by studying the model itself. In this paper we investigate a multiclass classification strategy designed to integrate feature selection, construction and model intelligibility goals into a distance-based classifier to improve its ability to build accurate and simple classifiers.

A well known learning method that implicitly conducts feature selection and construction is genetic programming (GP) [17], which has been proposed for classification [15, 7]. GP incorporates feature selection and construction by optimizing a population of programs constructed from a set of instructions that operate on the dataset features to produce a model. Compared to traditional ML approaches such as logistic regression and decision tree classification, GP makes fewer *a priori* assumptions about the data [22] and allows for various program representations [26]. In addition, GP has well-established methods for optimizing

*Corresponding author. Email: lacava@upenn.edu, Tel: +1 (413) 320-0544