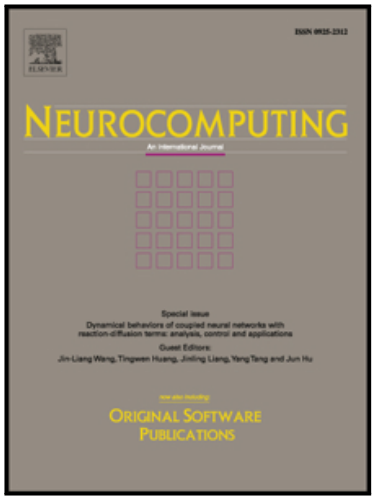
Communicated by Prof. H. Zhang

Accepted Manuscript



A Genetic Programming Approach for Feature Selection in Highly Dimensional Skewed Data

Felipe Viegas, Leonardo Rocha, Marcos Gonçalves, Fernando Mourão, Giovanni Sá, Thiago Salles, Guilherme Andrade, Isac Sandin

PII: S0925-2312(17)31471-6
DOI: 10.1016/j.neucom.2017.08.050
Reference: NEUCOM 18842

To appear in: *Neurocomputing*

Received date: 24 March 2016
Revised date: 23 August 2017
Accepted date: 30 August 2017

Please cite this article as: Felipe Viegas, Leonardo Rocha, Marcos Gonçalves, Fernando Mourão, Giovanni Sá, Thiago Salles, Guilherme Andrade, Isac Sandin, A Genetic Programming Approach for Feature Selection in Highly Dimensional Skewed Data, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.08.050

A Genetic Programming Approach for Feature Selection in Highly Dimensional Skewed Data

Felipe Viegas[b], Leonardo Rocha[a], Marcos Gonçalves[b], Fernando Mourão[a], Giovanni Sá[a], Thiago Salles[b], Guilherme Andrade[b], Isac Sandin[a]

[a]*Universidade Federal de São João Del Rei*
Department of Computer Science
São João Del Rei, MG, Brazil
{lcrocha,fhmourao,giovannisa,isac}@ufsj.edu.br
[b]*Universidade Federal de Minas Gerais*
Department of Computer Science
Belo Horizonte, MG, Brazil
{viegas,mgoncalv,tsalles,gnandrade}@dcc.ufmg.br

Abstract

High dimensionality, also known as the curse of dimensionality, is still a major challenge for automatic classification solutions. Accordingly, several feature selection (FS) strategies have been proposed for dimensionality reduction over the years. However, they potentially perform poorly in face of unbalanced data. In this work, we propose a novel feature selection strategy based on genetic programming, which is resilient to data skewness issues, in other words, it works well with both, balanced and unbalanced data. The proposed strategy aims at combining the most discriminative feature sets selected by distinct feature selection metrics in order to obtain a more effective and impartial set of the most discriminative features, departing from the hypothesis that distinct feature selection metrics produce different (and potentially complementary) feature space projections. We evaluated our proposal in biological and textual datasets. Our experimental results show that our proposed solution not only increases the efficiency of the learning process, reducing up to 83% the size of the data space, but also significantly increases its effectiveness in some scenarios.

Keywords: Feature Selection, Classification, Genetic Programming

1. Introduction

The organization and extraction of useful knowledge from the huge amount of data available in distinct applications is one of the biggest challenges in Computer Science nowadays. Machine learning (ML) techniques, such as Automatic Document Classification (ADC), have demonstrated to be a viable path towards facing such challenges. Particularly, ADC techniques aim at building effective models to associate documents with well-defined semantic categories in an automated way. ADC techniques are the core component of many important applications such as spam filtering [1], organization of topic directories [2], identification of writing styles or authorship [3], delayed chaotic systems [4], digital analysis [5, 6], among many others.

ADC methods usually exploit a supervised learning paradigm [7], which learns a model for classifying new unseen examples, given a set of previously labeled examples. More formally, given a training set $\mathbb{D}_{train} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, in which x_i is a feature vector that represents an example and y_i its class, the objective is to learn a classification