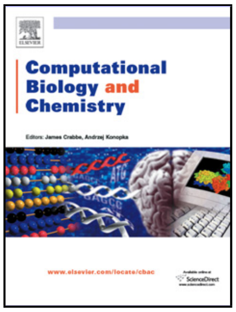

Title: CCFS: A Cooperating Coevolution Technique for Large
Scale Feature Selection on Microarray Datasets

Author: Mohammad K. Ebrahimpour Hossein
Nezamabadi-pour Mahdi Eftekhari

PII: S1476-9271(17)30031-2
DOI: https://doi.org/doi:10.1016/j.compbiolchem.2018.02.006
Reference: CBAC 6789

To appear in: *Computational Biology and Chemistry*

Received date: 17-1-2017
Revised date: 10-1-2018
Accepted date: 7-2-2018

Please cite this article as: Mohammad K. Ebrahimpour, Hossein Nezamabadi-pour, Mahdi Eftekhari, CCFS: A Cooperating Coevolution Technique for Large Scale Feature Selection on Microarray Datasets, *<![CDATA[Computational Biology and Chemistry]]>* (2018), https://doi.org/10.1016/j.compbiolchem.2018.02.006

CCFS: A Cooperating Coevolution Technique for Large Scale Feature Selection on Microarray Datasets

Mohammad K. Ebrahimpour*, Hossein Nezamabadi-pour**, and Mahdi Eftekhari*

*Computer Engineering Department
**Electrical Engineering Department
*,**Shahid Bahonar University of Kerman, Kerman, Iran

January 10, 2018

Abstract

Recently, advances in bioinformatics leads to microarray high dimensional datasets. These kind of datasets are still challenging for researchers in the area of machine learning since they suffer from small sample size and extremely large number of features. Therefore, feature selection is the problem of interest in the learning process in this area. In this paper, a novel feature selection method based on a global search (by using the main concepts of divide and conquer technique) which is called CCFS, is proposed. The proposed CCFS algorithm divides vertically (on features) the dataset by random manner and utilizes the fundamental concepts of cooperation coevolution by using a filter criterion in the fitness function in order to search the solution space via binary gravitational search algorithm. For determining the effectiveness of the proposed method some experiments are carried out on seven binary microarray high dimensional datasets. The obtained results are compared with nine state-of-the-art feature selection algorithms including Interact (INT), and Maximum Relevancy Minimum Redundancy (MRMR).The average outcomes of the results are analyzed by a statistical non-parametric test and it reveals that the proposed method has a meaningful difference to the others in terms of accuracy, sensitivity, specificity and number of selected features.

Keywords: Meta-heuristics, Cooperating Coevolving Feature Selection, Microarray Datasets, Divide and conquered algorithms.

1 Introduction

In the last two decades, handling DNA microarray high dimensional datasets (Garro et al. 2016, Mahajan et al. 2016) has created a new line of research in both machine learning (Ebrahimpour et al. 2017, Ebrahimpour & Eftekhari 2017) and bioinformatics (Chinnaswamy & Srinivasan 2016). These types of datasets suffer from small sample size and huge number of features since they measure gene expression (Bolón-Canedo, Sánchez-Marono, Alonso-Betanzos, Benítez & Herrera 2014*a*). Therefore, feature selection plays a crucial role in DNA microarray datasets, which removes the irrelevant and redundant features from the dataset (Chandrashekar & Sahin 2014, Xue et al. 2013). Thus, the learning algorithms concentrate on the important aspects of features that are useful for future predictions.

Generally speaking, feature selection algorithms can be divided into four main categories: filter, wrapper, hybrid and embedded approaches (Bolón-Canedo et al. 2013, Liu & Motoda 2012, Ebrahimpour & Eftekhari 2015*b*, 2018) that are investigated as follows:

Filter approaches: Filter methods undergo the feature selection process by considering the nature characteristics of each feature (Ebrahimpour & Eftekhari 2015*b*, Bolon-Canedo et al. 2011, Ebrahimpour & Eftekhari 2017, Ebrahimpour et al. 2017); for instance, they use statistical and mathematical characteristics of the features in feature selection process. Moreover, the classifier does not have any contribution in the decision making process. Consequently, filter approaches are fast and useful in high dimensional datasets;