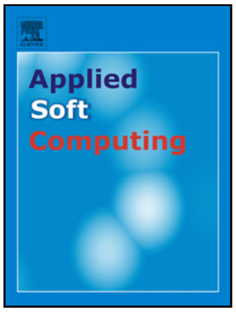

Title: Correlation Feature Selection based improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification

Author: Indu Jain Vinod Kumar Jain Renu Jain

PII: S1568-4946(17)30577-X
DOI: https://doi.org/doi:10.1016/j.asoc.2017.09.038
Reference: ASOC 4485

To appear in: *Applied Soft Computing*

Received date: 6-5-2016
Revised date: 27-8-2017
Accepted date: 26-9-2017

Please cite this article as: Indu Jain, Vinod Kumar Jain, Renu Jain, Correlation Feature Selection based improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification, *<![CDATA[Applied Soft Computing Journal]]>* (2017), https://doi.org/10.1016/j.asoc.2017.09.038

Correlation Feature Selection based improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification

Indu Jain[a], Vinod Kumar Jain[b,*], Renu Jain[a]

[a]*School of Mathematics and Allied Sciences (SOMAAS), Jiwaji University, Gwalior (M.P.), 474006, India*
[b]*PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Dumna Airport Road, P.O. Khamaria, Jabalpur (M.P.), India.*

Abstract

DNA microarray technology has emerged as a prospective tool for diagnosis of cancer and its classification. It provides better insights of many genetic mutations occurring within a cell associated with cancer. However, thousands of gene expressions measured for each biological sample using microarray pose a great challenge. Many statistical and machine learning methods have been applied to get most relevant genes prior to cancer classification. A two phase hybrid model for cancer classification is being proposed, integrating Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO). This model selects a low dimensional set of prognostic genes to classify biological samples of binary and multi class cancers using Naive-Bayes classifier with stratified 10-fold cross-validation. The proposed iBPSO also controls the problem of early convergence to the local optimum of traditional BPSO. The proposed model has been evaluated on 11 benchmark microarray datasets of different cancer types. Experimental results are compared with seven other well known methods, and our model exhibited better results in terms of classification accuracy and the number of selected genes in most cases. In particular, it achieved up to 100% classification accuracy for seven out of eleven datasets with a very small sized prognostic gene subset (up to $< 1.5\%$) for all eleven datasets.

Keywords: Microarray data analysis, cancer classification, improved binary particle swarm optimization (iBPSO), hybrid model, gene selection, naive-bayes.

1. Introduction

Parallel measurement of thousands of gene expressions using DNA microarray provides a big picture and better insights of many genetic alterations pertaining to cancer [1]. An early and accurate prognosis of cancer facilitates the proper line of treatment, and DNA microarray technology has shown great potential in diagnosis of cancer and its classification. The cancer datasets produced by microarray technology typically have thousands of gene expressions obtained from each biological sample. Furthermore, the number of samples are very less in comparison to gene expressions. These characteristics of cancer microarray data pose a great difficulty in analysis and its classification. However, only a few genes from these high dimensional datasets are significant for cancer classification [2]. The presence of redundant, irrelevant and noisy genes in the dataset degrades the computing efficiency as well as the classification accuracy of machine learning algorithms, particularly when samples are limited. Therefore, it becomes indispensable to alleviate irrelevant and redundant genes from the dataset using some feature selection methods [3]. Numerous methods are in use for feature (gene) selection which may be grouped into two categories: filters and wrappers. Filter method searches and evaluates either each gene individually (univariate filters) or the subset of genes

* Corresponding author. Tel.: +919425156298
Email addresses: indujain06@gmail.com (Indu Jain), vkjain@iiitdmj.ac.in (Vinod Kumar Jain), renujain3@rediffmail.com (Renu Jain)