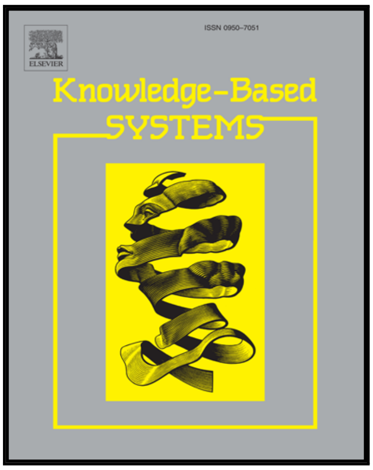

A Discrete Bacterial Algorithm for Feature Selection in Classification of Microarray Gene Expression Cancer Data

Hong Wang , Xingjian Jing , Ben Niu

PII: S0950-7051(17)30169-7
DOI: 10.1016/j.knosys.2017.04.004
Reference: KNOSYS 3884

To appear in: *Knowledge-Based Systems*

Received date: 10 November 2016
Revised date: 9 March 2017
Accepted date: 9 April 2017

Please cite this article as: Hong Wang , Xingjian Jing , Ben Niu , A Discrete Bacterial Algorithm for Feature Selection in Classification of Microarray Gene Expression Cancer Data, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.04.004

A Discrete Bacterial Algorithm for Feature Selection in Classification of Microarray Gene Expression Cancer Data

Hong Wang [a, b], Xingjian Jing [a], Ben Niu[b,c*]

[a] Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong

[b] Shenzhen University, Shenzhen, P.R. China

[c] School of Computing, Information, and Decision Systems Engineering, Arizona State University, USA

Abstract: When mining in high dimensional data, the curse of dimensionality is one of the major difficulty to overcome. In this paper, a weighted feature selection strategy is developed and embedded in bacterial based algorithms to reduce the feature dimension in classification. The proposed weighted feature selection strategy distinguishes the features by their classification performances as well as the occurrence frequency in population according to the two matrices. The objectives of minimizing the number of features, maximizing the performance, and minimizing the computational cost are all considered. Regarding the drawback of bacterial based algorithms, bacterial colony optimization based feature selection algorithm is proposed to decrease the computational complexity as well as improve the search ability even in discrete optimization problems. To test the effectiveness of the proposed feature selection method, four bacterial based methods with the weighted strategy embedded have been compared with four classical feature selection methods and three well-known population based algorithms using 15 cancer micro-array datasets with different numbers of features and classes. The results show that the weighted feature selection strategies embedded have improved the feature selection capability of bacterial algorithms. The new proposed mechanisms embedded in bacterial colony optimization method can overcome the limitation of the traditional bacterial based algorithms using premature termination to decrease the computational time, and provide comparable or in most cases better solutions than other feature selection methods considered in the comparison.

Key Words: Feature selection, bacterial colony optimization, bacterial foraging optimization, cancer classification

1 Introduction

In the classification of microarray gene expression cancer problems, the high dimensional features often brings the non-inconsiderable barrier for researchers. Actually, in classification tasks, the unlabeled instances are required to be classified into specific groups according to informative features. The reality is that not all features are useful for classification, and some redundant and irrelevant features may even serve as obstacles. Consequently, the accuracy of classification is also decreased [1]. The challenge is that, without previous knowledge, it is difficult to distinguish the most representative and compact features from the useless ones when the dimensional space is rather high. Confronting with such high dimensional feature characterization problem, feature selection (FS), as the effective tool, have been wildly applied to select a relative small but useful feature subsets from available features [2].

Feature selection in classification is regarded as the optimization problem with the main objective to maximize the classification accuracy rate with the smaller size of features. Also, many researchers treat this problem as a multi-objective optimization problem with the additional objectives like minimizing the number of selected features, minimizing the computational cost (evaluated by computational time) [3], etc. Even so, it does not mean that the smaller number of features guarantee the better solutions. The purpose of feature selection is to eliminate the redundant and irrelevant features in great deal to achieve the highest classification accuracy. Thus, all in all, the classification performance is the most important objective it needs to achieve, and the computational cost and the number feature are minor objectives to be considered in addressing the classification problems. While once the algorithm is used as the on-line analyzer, and the computational cost as well as classification accuracy are both important. In this situation, the principle objective we may want to accomplish is to achieve the high quality classification in a limited computational time regardless the number of features. Therefore, in this paper, feature selection is regarded as a single objective optimization problem with the classification accuracy as the kernel objective to select the feature sets with the pre-specified optimal number of features. Besides, the developed feature selection algorithms need to accomplish selection procedure at the accepted time domain to achieve the efficiency using compacted feature sets.

Inspired by chemotactic (foraging) behavior of E. coli bacteria, Bacterial Foraging Optimization (BFO) proposed by Passino[4] and bacteria chemotaxis (BC) presented by Müller [5] are two earliest bacterial based algorithms for optimization problems. Currently, these two bacterial based algorithms start a new heuristic family in computational intelligence due to its global searching capability for control and optimization. Since now, bacterial based FS methods developed, mostly, are cooperative with other methods. Cho combined the BFO with mutual information for feature selection in classification [3]. In [6], agent genetic algorithm based on bacteria foraging strategy (BFOA-L) was proposed to integrate with the neural network to implement the fuzzy logic reasoning in feature selection. Also, BFO working

[*] Corresponding author.
E-mail addresses: drniube@gmail.com (B. Niu)