Available online at www.sciencedirect.com
ScienceDirect
Journal of Bionic Engineering 10 (2013) 231–241

An Ant Colony Optimization Based Dimension Reduction Method for High-Dimensional Datasets

Ying Li[1,2], Gang Wang[1,2,3], Huiling Chen[4], Lian Shi[1,2], Lei Qin[1,2]

1. *College of Computer Science and Technology*, *Jilin University*, *Changchun* 130012, *P. R. China*
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education*, *Jilin University*, *Changchun* 130012, *P. R. China*
3. *College of Geo Exploration Science and Technology*, *Jilin University*, *Changchun* 130026, *P. R. China*
4. *College of physics and electronic information*, *Wenzhou University*, *Wenzhou* 325035, *P. R. China*

Abstract

In this paper, a bionic optimization algorithm based dimension reduction method named Ant Colony Optimization -Selection (ACO-S) is proposed for high-dimensional datasets. Because microarray datasets comprise tens of thousands of features (genes), they are usually used to test the dimension reduction techniques. ACO-S consists of two stages in which two well-known ACO algorithms, namely ant system and ant colony system, are utilized to seek for genes, respectively. In the first stage, a modified ant system is used to filter the nonsignificant genes from high-dimensional space, and a number of promising genes are reserved in the next step. In the second stage, an improved ant colony system is applied to gene selection. In order to enhance the search ability of ACOs, we propose a method for calculating priori available heuristic information and design a fuzzy logic controller to dynamically adjust the number of ants in ant colony system. Furthermore, we devise another fuzzy logic controller to tune the parameter (q_0) in ant colony system. We evaluate the performance of ACO-S on five microarray datasets, which have dimensions varying from 7129 to 12000. We also compare the performance of ACO-S with the results obtained from four existing well-known bionic optimization algorithms. The comparison results show that ACO-S has a notable ability to generate a gene subset with the smallest size and salient features while yielding high classification accuracy. The comparative results generated by ACO-S adopting different classifiers are also given. The proposed method is shown to be a promising and effective tool for mining high-dimension data and mobile robot navigation.

Keywords: gene selection, feature selection, ant colony optimization, high-dimensional data

doi: 10.1016/S1672-6529(13)60219-X

1 Introduction

The advent of DNA microarray technology has provided not only the ability to measure the expression levels of thousands of genes simultaneously in a single experiment but also the possibility to identify diagnosis disease[1]. Therefore, an overall understanding of the cell can be obtained. The gene expression data is very different from any of the data. First, it has a very high dimensionality, usually contains thousands to tens of thousands of genes. Second, publicly available data size is very small. Third, most genes are irrelevant to cancer distinction. As a result, existing classification methods turn out to be not efficient and effective to handle this kind of data[2–3]. The irrelevant gene expression data leads to a high computational complexity and makes it impossible to discover relevant genes. The reason of performing gene selection prior to cancer classification is twofold. One is that performing gene selection can help reduce data size, and thus cutting down the running time. The other and more important one is that gene selection can eliminate a great number of irrelevant genes so as to improve the classification accuracy[4–5].

With the proliferation of high-dimensional data, Feature Selection (FS) has become an indispensable task of a learning process. FS aims to select a good subset of features from the original set of features without losing a suitably high accuracy in representing the original features, in which there exists abundance of noise, spurious information, and irrelevant and redundant features[6–7].

Corresponding author: Gang Wang
E-mail: wanggang.jlu@gmail.com

FS helps to improve the quality and speed of learning algorithms and to enhance the comprehensibility of the constructed models by removing irrelevant and redundant features. Generally, the universal algorithms of FS are often classified into three modalities, wrapper, filters and embeded. Accordingly, FS has made a strong impact on many fields, including gene selection, pattern recognition, data mining, image mining, and text categorization[8–10].

In order to deal with these particular characteristics of microarray data, the obvious need for dimension reduction techniques was realized, and soon their applications became a de facto standard in the field. In 2001 the microarray analysis was still claimed to be in its infancy. Since then a considerable and valuable efforts have been done to contribute new and adaptive known FS methodologies. Many filter approaches have been proposed (t-statistics, χ^2-statistics, informative gain, signal-noise ratio, Pearson correlation coefficient and combination of several feature filtering algorithms). Some wrapper-based approaches have been provided and widely applied in bioinformatics, such as Genetic Algorithm (GA)[11], Particle Swarm Optimization (PSO)[12], Ant Colony Optimization (ACO)[13,14,15] and Simulating Annealing (SA)[16]. An increasing number of researches make use of the embedded capacity of several classifiers to discard input features. Variations of the popular method originally proposed for gene expression domains by Guyon *et al*.[17], use the weights of the variables in the Support Vector Machine (SVM) formulation to discard features with small weights. These methods have been broadly and successfully applied in the Mass Spectrometry (MS) domain[18–20]. Based on a similar framework, the weights of the input masses in a neural network classifier have been used to rank the features' importance by Ball *et al*.[21]. The embedded capacity of random forests[22] and other types of decision tree based algorithms[23] constitutes an alternative embedded FS strategy. Even though these approaches have obtained prominent performance in gene expression data analysis, some congenital drawbacks still make these approaches unsatisfying. On the one hand, it is hard to search the large range due to the convergent ability for high-dimensional data. On the other hand, it is difficult to give a powerful evaluation criterion owing to limited knowledge about the feature mutual relation.

Several state-of-the-art methods of solving the gene selection problems via ACO have been developed. Shi *et al*.[24] proposed an ACO and Rough Set (RS) based ensemble approach. ACO and RS theory are incorporated to select a subset of all the trained component classifiers for aggregation. Wu *et al*.[25] used filter method to rank the genes in terms of their expression difference, and then selected 'important' genes with high 'score'. An ACO is used in clustering gene expression data, and SVM is applied to validate the classification performance of candidate genes. Robins *et al*.[26] modified ACO to apply into several high-dimensional data sets. Patil *et al*.[27] proposed an ACO/random forest based hybrid filter-wrapper search technique, which traverses the search space and selects a feature subset with high classifying ability. Ke *et al*.[28] introduced a new approach based on ACO for the purpose of attribute reduction. Numerical experiments were carried out on thirteen small or medium-sized datasets and three gene expression datasets. Yu *et al*.[29] modified ACO to select tumor-related marker genes, and used SVM as a classifier to evaluate the performance of the extracted gene subset.

For processing high-dimensional data, the ACO based research mentioned above paid no attention to the dynamic performance of ACO, thus important genes were lost during the search progress. Moreover, few effective methods have been designed to exactly evaluate the priori available heuristic information for ants. In this respect, we aim at studying the intrinsic dynamics and the feature importance evaluation in ACO. To this end, Fuzzy logic control[30–31] is used to simulate the dynamic search for the swarm, and Information Gain (IG) and F-score[32] are used to give an accurate estimation for the priori available heuristic information. We propose an ACO based framework for gene selection, which consists of two stages. In the first stage, the pivotal genes are chosen by modified ant system while the most important genes are selected by the modified ant colony system in the second stage. In order to certify the performance of the improved ACOs, we compared the two stages with only stage II. From the results, we could see that good results are obtained by the second stage only, which surpass the other algorithms, and better results are obtained by the two stage selection method.

The main contributions of this paper are described as follows.

(1) We propose a two-stage framework for gene selection so that the modified ant system and improved