Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 32 (2008) 53–60
www.elsevier.com/locate/compbiolchem

Research Article

Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data

Qi Shen, Wei-Min Shi *, Wei Kong

Chemistry Department, Zhengzhou University, Zhengzhou 450052, China

Received 7 January 2007; received in revised form 10 October 2007; accepted 14 October 2007

Abstract

Gene expression data are characterized by thousands even tens of thousands of measured genes on only a few tissue samples. This can lead either to possible overfitting and dimensional curse or even to a complete failure in analysis of microarray data. Gene selection is an important component for gene expression-based tumor classification systems. In this paper, we develop a hybrid particle swarm optimization (PSO) and tabu search (HPSOTS) approach for gene selection for tumor classification. The incorporation of tabu search (TS) as a local improvement procedure enables the algorithm HPSOTS to overleap local optima and show satisfactory performance. The proposed approach is applied to three different microarray data sets. Moreover, we compare the performance of HPSOTS on these datasets to that of stepwise selection, the pure TS and PSO algorithm. It has been demonstrated that the HPSOTS is a useful tool for gene selection and mining high dimension data.

Keywords: Particle swarm optimization; Tabu search; Gene selection; Gene expression data

1. Introduction

High-density DNA microarrays are one of the most powerful tools for functional genomic studies and the development of microarray technology allows for measuring expression levels of thousands of genes simultaneously Schena et al. (1995). Recent studies have shown that one of the most important applications of microarrays is tumor classification (Cho et al., 2003; Li et al., 2004). Gene selection is an important component for gene expression-based tumor classification systems. Microarray experiments generate large datasets with expression values for thousands even tens of thousands of genes but not more than a few tissue samples. Most of the genes monitored in microarray may be irrelevant to analysis and the use of all the genes may potentially inhibit the prediction performance of classification rule by masking the contribution of the relevant genes (Li, 2006; Li and Yang, 2002; Stephanopoulos et al., 2002; Nguyen and Rocke, 2002; Biceiato et al., 2003; Tan et al., 2004). An efficient way to solve this problem is gene selection and the selection of discriminatory genes is critical to improving the accuracy and decrease computational complexity and cost. By selecting relevant genes, conventional classification techniques can be applied to the microarray data. Gene selection may highlight those relevant genes and it could enable biologists to gain significant insight into the genetic nature of the disease and the mechanisms responsible for it (Guyon et al., 2002; Wang et al., 2005).

Several gene selections techniques have been employed in classification problems, such as *t*-test filtering approach, as well as some artificial intelligence techniques such as genetic algorithms (GAs), evolution algorithms (EAs) (Golub et al., 1999; Furey et al., 2000; Xiong et al., 2001; Peng et al., 2003; Li et al., 2005; Tibshirani et al., 2002; Sima and Dougherty, 2006), simulated annealing, tabu search and particle swarm optimization.

Particle swarm optimization (PSO) algorithm (Kennedy and Eberhart, 1995; Shi and Eberhart, 1998; Clerc and Kennedy, 2002) is a recently proposed algorithm by James Kennedy and R.C. Eberhart in 1995, motivated by social behavior of organisms such as bird flocking and fish schooling. Particle swarm optimization comprises a very simple concept, and can be implemented in a few lines of computer code. It requires only few parameters to adjust, and is computationally inexpensive in

* Corresponding author. Tel.: +86 371 67767957; fax: +86 371 67763220.
E-mail address: shiweimin@zzu.edu.cn (W.-M. Shi).

1476-9271/$ – see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.compbiolchem.2007.10.001

terms of both memory requirements and speed. A modified discrete PSO algorithm has been proposed in our previous study (Shen et al., 2004a,b, in press) to reduce dimension and shown satisfied performance. Although PSO has proved to be a potent search technique for solving optimization, there are still many complex situations where the PSO tends to converge to local optima and does not perform particularly well.

Tabu search (TS) is a powerful optimization procedure that has been successfully applied to a number of combinatorial optimization problems Glover (1986). It has the ability to avoid convergence to local minima by employing a flexible memory system. But the convergence speed of TS depends on the initial solution and the parallelism of PSO population would help the TS find the promising regions of the search space very quickly.

In this paper, we develop a hybrid PSO and TS (HPSOTS) approach for gene selection for tumor classification. The incorporation of TS as a local improvement procedure enables the algorithm HPSOTS to overleap local optima and show satisfactory performance. The formulation and corresponding programming flow chart are presented in details in the paper. To evaluate the performance of HPSOTS, the proposed approach is applied to three publicly available microarray datasets. Moreover, we compare the performance of HPSOTS on these datasets to that of stepwise selection, the pure TS and PSO algorithm. It has been demonstrated that the HPSOTS is a useful tool for gene selection and mining high dimension data.

2. Methods

2.1. Modified particle swarm optimization

PSO Kennedy and Eberhart (1995), Shi and Eberhart (1998) and Clerc and Kennedy (2002) is a stochastic global optimization technique and can be used for gene selection. PSO carries out a search based on population of individuals or particles and each particle represents a potential solution in the search space. During flight, each particle with a velocity which is adjusted its position according to its own experience and the other particle's experience until stopping criteria is satisfied. Suppose that the problem space is D-dimensional, then the position of the ith particle is represented as $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$. Velocity, the rate of the position change for particle i is represented as $v_i = (v_{i1}, v_{i2}, \ldots v_{iD})$. The best previous position of the ith particle that gives the best fitness value is expressed as $p_i = (p_{i1}, p_{i2}, \ldots, p_{iD})$. The position of the best particle in the swarm is denoted as $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$. For a discrete problem expressed in a binary notation, a particle moves in a search space restricted to 0 or 1 on each dimension. In binary problem, updating a particle represents changes of a bit that should be in either state 1 or 0 and the velocity represents the probability of bit x_{id} taking the value 1 or 0. In every iteration, each particle is updated by following the two best values.

According to information sharing mechanism of PSO, a modified discrete PSO (Shen et al., 2004a,b) was proposed as follows. The velocity v_{id} of every individual is a random number in the range of (0,1). The resulting change in position then is defined by the following rule:

$$\text{If } (0 < v_{id} \le a), \quad \text{then } x_{id}(\text{new}) = x_{id}(\text{old}) \tag{1}$$

$$\text{If } \left(a < v_{id} \le \frac{(1+a)}{2}\right), \quad \text{then } x_{id}(\text{new}) = p_{id} \tag{2}$$

$$\text{If } \left(\frac{(1+a)}{2} < v_{id} \le 1\right), \quad \text{then } x_{id}(\text{new}) = p_{gd} \tag{3}$$

where a is a random value in the range of (0,1) named static probability. In this study static probability a equals to 0.5. Though the velocity in the modified discrete PSO is different from that in continuous version of PSO, information sharing mechanism and updating model of particle by following the two best positions is the same in two PSO versions. The details of modified PSO have been described elsewhere (Shen et al., 2004a,b).

2.2. Tabu search

Tabu search (TS) was invented by Glover (1986) and has been used to solve a wide range of hard optimization problems. TS is an iterative procedure designed for the solution of optimization problems. TS starts with a random solution and evaluate the fitness function for the given solution. Then all possible neighbors of the given solution are generated and evaluated. A neighbor is a solution which can be reached from the current solution by a simple, basic transformation. If the best of these neighbors is not in tabu list then pick it to be the new current solution. The tabu list keeps track of previously explored solutions and prohibits TS from revisiting them again. Thus, if the best neighbor solution is worse than the current design, TS will go uphill. In this way, local minima can be overcome. Any reversal of these solutions or moves is then forbad move and is classified as tabu. Some aspiration criteria which allow overriding of tabu status can be introduced if that moves is still found to lead to a better fitness with respect to the fitness of the current optimum. If no more neighbors are present (all are tabu), or when during a predetermined number of iterations no improvements are found, the algorithm stops. Otherwise, the algorithm continues the TS procedures.

2.3. Classification by hybrid PSO and TS (HPSOTS) approach

Although PSO has the advantages of good convergent property and is effective on solving optimization, after some generations the population diversity would be greatly reduced and the PSO algorithm might lead to a premature convergence to a local optimum. TS is a powerful stochastic optimization technique, which can theoretically converge asymptotically to a global optimum solution, but it will take much time to reach the near-global minimum. The incorporation of TS into PSO as a local improvement procedure enables the algorithm to maintain the population diversity and prevent leading to misleading local optima. In the present work strings of binary bits were adopted to code all the particles of the modified discrete PSO. Each binary bits coded string (particle) stands for a set of genes, which are