A Postprint of an APSO Algorithm Combining Relief-F and Decision Trees for Gene Selection

Authors: Ye Chaochao, Pan Julong

Date: 2018-05-20T00:00:00+00:00

Abstract

Due to the high dimensionality, high noise, and small sample size characteristics of gene expression data, gene selection has always been a major challenge in tumor classification. To improve the accuracy of tumor classification while ensuring the efficiency of gene selection, this study proposes an adaptive particle swarm optimization (APSO) algorithm (R-C-APSO) that combines Relief-F and CART decision tree. This method first utilizes Relief-F to rapidly filter out a large number of irrelevant genes and noise, thereby narrowing the scope of gene selection; subsequently, it employs the CART decision tree as the fitness function and uses the APSO algorithm to conduct the final search for genes. Through experimental analysis on six datasets, the results demonstrate that R-C-APSO achieves high classification accuracy and fast gene selection speed, and exhibits good stability.

Full Text

A Novel APSO Algorithm for Gene Selection Combining Relief-F and Decision Tree

Ye Chaochao, Pan Julong†
(College of Information Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract

Gene selection has long been a major challenge in tumor classification due to the high dimensionality, high noise, and small sample size characteristics of gene expression data. To improve tumor classification accuracy while ensuring gene selection efficiency, this paper proposes an Adaptive Particle Swarm Optimization (APSO) algorithm that integrates Relief-F and CART decision tree, designated as R-C-APSO. The method first employs Relief-F to rapidly filter out

large numbers of irrelevant genes and noise, thereby narrowing the search space for gene selection. It then utilizes the CART decision tree as the fitness function, with the APSO algorithm performing the final gene search. Experimental analysis on six datasets demonstrates that R-C-APSO achieves high classification accuracy with fast gene selection speed and exhibits good stability.

Keywords: gene selection; adaptive particle swarm; decision tree; tumor classification

0 Introduction

The rapid development of DNA microarray technology enables large-scale gene expression data measurement in a single experiment. Gene expression data is characterized by high dimensionality, high noise, and small sample size [1]. Among thousands of genes, only a small fraction plays a role in tumor diagnosis. Using all feature genes directly for tumor classification would make computation complex and inefficient. Therefore, gene selection is a crucial step in data preprocessing for tumor classification, as it can improve classification accuracy by removing irrelevant genes while reducing computational cost [2].

Gene selection is one of the most important tasks in microarray expression data analysis because it helps discover disease mechanisms, reduce clinical diagnosis costs, and improve tumor classification accuracy [3]. Filter methods, as one category of feature selection approaches, are widely adopted due to their simplicity and computational efficiency. These methods typically use statistical techniques to select features independently of the classification model, commonly including mutual information, t-statistics, and information gain. However, existing methods have limitations: two-sample t-statistic can only handle binary classification problems [4]; the two-stage feature selection method based on mutual information and genetic algorithm [5] and the gene selection method based on information gain and improved simplified particle swarm optimization [6] often employ discretization when processing continuous feature values, which may lead to loss of important feature information. While Relief has been introduced into SNP genome-wide association analysis to eliminate large numbers of irrelevant genes and achieve high classification accuracy [7], it is limited to binary problems. In 1994, Kononenko extended Relief to Relief-F [8], which can handle multi-class problems and continuous feature values, making it an efficient data preprocessing algorithm.

Wrapper methods represent another important category of feature selection approaches, comprising a learner and a search algorithm. Notable search algorithms include Genetic Algorithms and Particle Swarm Optimization (PSO) [9]. Proposed by James Kennedy and Russell Eberhart in 1995, PSO was inspired by bird flocking behavior and has been widely applied due to its few parameters, simple implementation, and good global search capability [10-13]. Previous studies have demonstrated PSO' s effectiveness in gene selection: a novel local search strategy with PSO for gene subset selection [14]; basic PSO combined

with decision tree classifiers for tumor classification [15]; support vector machine combined with improved binary PSO for near-optimal feature subset selection [16]; and a hybrid feature selection method combining correlation coefficients with PSO that achieved high classification accuracy on three tumor datasets [17]. These studies collectively demonstrate that PSO is an excellent algorithm for gene selection. However, when feature dimensionality is extremely high and the search space contains multiple local optima, PSO suffers from premature convergence, local optimum trapping, and high computational cost.

Based on this analysis, this paper proposes a novel and efficient gene selection algorithm called R-C-APSO. On one hand, it leverages Relief-F' s fast filtering capability for preliminary gene selection, limiting the number of genes to accelerate the subsequent search algorithm. On the other hand, it introduces adaptive inertia weights into PSO, employing the Adaptive Particle Swarm Optimization (APSO) algorithm with stronger global search capability for final gene selection. Additionally, this work selects Classification and Regression Tree (CART) as the fitness function for APSO. Research on combining CART with APSO remains limited. Since CART inherently possesses feature selection capability, using it as the fitness function to validate gene selection effectiveness will further improve tumor classification accuracy.

1.2 Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization (PSO) is a heuristic global optimization method that achieves approximate optimal solutions in multi-dimensional spaces by simulating the social behavior of bird flocks [18]. Each object in PSO is called a particle, with each particle having position and velocity attributes. Assuming the algorithm operates in n-dimensional space, the position and velocity of particle i at iteration p can be represented as:

$$\mathbf{x}_i^p = [x_{i1}^p, x_{i2}^p, ..., x_{in}^p]$$
$$\mathbf{v}_i^p = [v_{i1}^p, v_{i2}^p, ..., v_{in}^p]$$

During iteration, each particle evaluates its current position based on the fitness function. Let $pbest_{id}^p$ denote the best position of particle i in the first p iterations along dimension d, and $gbest_d^p$ denote the best position of the entire particle swarm in the first p iterations along dimension d. The particle determines its next position and velocity through its current position, velocity, and these two best positions. The update formula for particle i is:

$$v_{id}^{p+1} = \omega v_{id}^p + c_1 r_1 (pbest_{id}^p - x_{id}^p) + c_2 r_2 (gbest_d^p - x_{id}^p) \tag{1}$$

$$x_{id}^{p+1} = \begin{cases} 1, & \text{if sigmoid}(v_{id}^{p+1}) > U(0,1) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where c_1 and c_2 are learning coefficients; r_1 and r_2 are random numbers between 0 and 1; ω is the inertia weight during iteration, typically a constant; and $U(0,1)$ represents a random number uniformly distributed between 0 and 1. Additionally, each particle's velocity during iteration must be constrained within $[V_{min}, V_{max}]$ [9,15].

2 R-C-APSO Algorithm Combining Relief-F and Decision Tree

2.1 Adaptive Particle Swarm Optimization (APSO) Algorithm

Although PSO possesses strong search capabilities, it often falls into local optima when facing multi-modal problems, missing the global optimum [13]. To enhance the overall search capability of the swarm, this paper introduces an adaptive inertia weight ω_p to update particle velocities:

$$v_{id}^{p+1} = \omega_p v_{id}^p + c_1 r_1 (pbest_{id}^p - x_{id}^p) + c_2 r_2 (gbest_d^p - x_{id}^p) \tag{3}$$

The adaptive inertia weight is calculated as:

$$\omega_p = (\omega_{max} - \omega_{min}) \times \exp\left(-\eta \times \frac{p}{P_{max}}\right) + \omega_{min} \tag{4}$$

where ω_{max} and ω_{min} are the maximum and minimum inertia weights, respectively; η is an empirical value in [20,55]; p is the current iteration number; and P_{max} is the preset maximum iteration count. Equation (4) shows that ω_p decreases exponentially during iteration. The inertia weight balances global and local search capabilities in PSO [13]. Larger ω_p values enhance APSO' s global search ability, while smaller values strengthen local search.

2.2 Fitness Function

The fitness function serves as the metric for evaluating particle quality, significantly impacting APSO' s convergence speed and ability to find optimal solutions. This paper employs the CART decision tree as the fitness function because its inherent feature selection capability provides better validation of feature subsets and better evaluation of particle quality.

Assume a sample set D containing n classes, where the proportion of class m samples is p_m $(m = 1, 2, ..., n)$. The purity of D can be measured by:

$$\text{Gini}(D) = \sum_{m=1}^{n} p_m(1 - p_m) = 1 - \sum_{m=1}^{n} p_m^2 \tag{5}$$

The Gini value Gini(D) is negatively correlated with purity; a smaller value indicates higher purity. When CART selects splitting attributes, it uses the

Gini index [20]. Suppose attribute b splits sample set D into F branches, with branch f' s sample set denoted as D_f. The Gini index for attribute b is defined as:

$$\text{GiniIndex}(b) = \sum_{f=1}^{F} \frac{|D_f|}{|D|} \text{Gini}(D_f) \tag{6}$$

where $|D|$ and $|D_f|$ represent the sample counts in sets D and D_f, respectively. CART selects the attribute with the minimum Gini index as the splitting attribute. A smaller Gini index means higher sample purity in the resulting branch sets.

2.3 Algorithm Description

To improve tumor classification accuracy while ensuring gene selection efficiency, this paper proposes a novel gene selection algorithm: R-C-APSO (Relief-F and CART decision tree-based Adaptive Particle Swarm Optimization for gene selection). This method fully leverages Relief-F' s filtering capability, APSO' s nonlinear search ability, and CART' s inherent feature selection advantage.

First, the algorithm uses Relief-F to compute each gene' s relevance statistic according to equation (1) and ranks genes by these values in descending order. A larger relevance statistic indicates greater importance for tumor classification. Next, it filters the ranked genes, retaining only the top T genes while removing large numbers of irrelevant genes and noise. Finally, it initializes the APSO population using these selected genes and employs the powerful APSO algorithm to select the final gene subset. Since Relief-F significantly reduces gene dimensionality in this preliminary stage, it saves substantial time for APSO' s search process.

During APSO population initialization, particle i' s initial velocity and position along dimension d are calculated through equations (7) and (8), where $U(0,1)$ represents a random number uniformly distributed between 0 and 1. During the APSO search phase, the CART decision tree classifier serves as the fitness function to evaluate each searched gene subset based on classification accuracy. CART' s feature selection capability and fast model construction make it an effective fitness function for evaluating gene subsets, further enhancing tumor classification accuracy and gene selection efficiency.

The R-C-APSO algorithm implementation proceeds as follows:

a) Perform zero-mean standardization on the original features of each sample;

b) Use Relief-F to compute each gene' s relevance statistic according to equation (1), where the repeated sampling count M equals the sample size, ensuring each sample is selected on average;

c) Rank all genes by relevance statistic and retain only the top T genes (here T = 100);

d) Set APSO population size N and maximum iteration count P_{max}, then initialize the swarm (N = 20, P_{max} = 100);

e) Set APSO learning coefficients $c_1 = c_2 = 2$; particle maximum velocity $V_{max} = 4$, minimum velocity $V_{min} = -4$; maximum inertia weight $\omega_{max} = 0.9$ and minimum inertia weight $\omega_{min} = 0.4$ according to reference [21];

f) Calculate adaptive inertia weight using equation (4), then compute each particle' s fitness value using the CART decision tree classifier;

g) Update the swarm' s global best position and each particle' s personal best position based on fitness values, then update particle velocities and positions using equations (3) and (2);

h) Check if maximum iteration count is reached; if not, return to step f); otherwise, terminate.

3 Experimental Results and Analysis

To validate R-C-APSO' s effectiveness, experiments were conducted on six publicly available gene expression datasets: SRBCT, Brain_{Tumor1}, 9_{Tumors}, Prostate_{Tumor}, Brain_{Tumor2}, and DLBCL. Dataset details are shown in . Experiments were performed on a 64-bit Windows 7 system with Python 2.7, an i5-3230M@2.6 GHz processor, and 8 GB RAM.

Using CART decision tree as the classifier, each dataset underwent 10 experimental runs with five-fold cross-validation. The average of these 10 runs served as the final classification accuracy. To comprehensively evaluate R-C-APSO' s performance, comparative experiments were conducted against: CART algorithm, Relief-F algorithm, PSO algorithm with CART as fitness function (CART-PSO), and SVM algorithm.

presents the experimental results for all six datasets across the five algorithms. The results show that R-C-APSO achieves the highest classification accuracy on all datasets. Compared with classic SVM, CART-PSO demonstrates superior performance on most datasets except Brain_{Tumors} and DLBCL where accuracies are similar, indicating that combining CART decision tree with PSO is already an effective gene selection method. Furthermore, comparing rows 2 and 3 in reveals that CART algorithm's classification accuracy improves significantly after Relief-F processing, demonstrating Relief-F' s suitability for preliminary gene filtering and its ability to remove large numbers of irrelevant genes and noise.

CART-PSO and R-C-APSO represent the two highest-accuracy algorithms among the five, both being PSO-based. To further compare these algorithms, CPU runtime and iteration processes were recorded, as shown in and [Figure 1: see original paper]-[Figure 3: see original paper].

clearly shows that R-C-APSO' s CPU runtime is substantially less than CART-PSO across all datasets. In adaptive particle swarm optimization, fitness function complexity constitutes the main component of algorithmic complexity. Since this work uses CART decision tree as the fitness function with complexity increasing with feature dimensionality, Relief-F' s dimensionality reduction effectively decreases APSO' s complexity, greatly improving R-C-APSO' s time efficiency.

Moreover, [Figure 1: see original paper]-[Figure 3: see original paper] clearly demonstrate that R-C-APSO has better search starting points than CART-PSO across all six datasets and maintains higher classification accuracy throughout subsequent iterations. Under conditions achieving equivalent classification accuracy, R-C-APSO requires less time and fewer iterations, offering greater advantages in resource-limited or time-constrained scenarios.

In summary, R-C-APSO outperforms the other four algorithms in classification accuracy while providing faster gene selection speed and better stability compared to CART-PSO.

4 Conclusion

This paper fully leverages the nonlinear search capability of adaptive particle swarm optimization, the filtering performance of Relief-F, and the splitting advantage of CART decision tree to propose R-C-APSO for gene selection. The algorithm not only improves tumor classification accuracy but also significantly reduces computational time compared to particle swarm optimization algorithms, demonstrating good stability. Experimental analysis confirms that R-C-APSO is an effective gene selection algorithm.

Regarding parameter selection for the adaptive particle swarm optimization component, values such as maximum and minimum inertia weights were primarily referenced from previous literature. Future work will involve additional research and experiments to optimize these parameters for improved convergence speed and global search capability. Furthermore, while this work evaluates gene selection methods primarily based on final tumor classification performance, in-depth analysis of the number and characteristics of selected genes remains as future work.

References

[1] Ghosh S, Mitra S, Dattagupta R. Fuzzy clustering with biological knowledge for gene selection [J]. Applied Soft Computing, 2014, 16: 102-111.
[2] Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization

and genetic algorithm [J]. Soft Computing, 2008, 12 (11): 1039-1048.
[3] Chen Y, Wang L, Li L, et al. Informative gene selection and the direct classification of tumors based on relative simplicity [J]. BmcBioinformatics, 2016, 17 (1): 1-16.
[4] Varma S, Simon R. Iterative class discovery and feature selection using Minimal Spanning Trees [J]. Bmc Bioinformatics, 2004, 5 (1): 126-126.
[5] 裘国永, 王娜, 汪万紫. 基于互信息和遗传算法的两阶段特征选择方法 [J]. 计算机应用研究, 2012, 29 (8): 2903-2905.
[6] Lai C M, Yeh W C, Chang C Y. Gene selection using information gain and improved simplified swarm optimization [J]. Neurocomputing, 2016, 218: 331-438.
[7] 吴红霞, 吴悦, 刘宗田, 等. 基于 Relief 和 SVM-RFE 的组合式 SNP 特征选择 [J]. 计算机应用研究, 2012, 29 (6): 2074-2077.
[8] Kononenko I. Estimating attributes: analysis and extensions of RELIEF [C]// Proc of European Conference on Machine Learning on Machine Learning. New York: Springer-Verlag, 1994: 171-182.
[9] Zhang Y, Wang S, Ji G. A comprehensive survey on particle swarm optimization algorithm and its applications [J]. MathematicalProblems in Engineering, 2015, 2015 (1): 1-38.
[10] Ghamisi P, Benediktsson J A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization [J]. IEEE Geoscience & Remote Sensing Letters, 2014, 12 (2): 309-313.
[11] Xi M, Sun J, Li L, et al. Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine [J]. Computational & Mathematical Methods in Medicine, 2016, 2016 (9): 1-9.
[12] Teng X, Dong H, Zhou X. Adaptive feature selection using v-shaped binary particle swarm optimization [J]. Plos One, 2017, 12 (3).
[13] Lu H, Du B, Liu J, et al. A kernel extreme learning machine algorithm based on improved particle swam optimization [J]. Memetic Computing, 2017, 9 (2): 121-128.
[14] Moradi P, Gholampour M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy [J]. Applied Soft Computing, 2016, 43: 117-130.
[15] Chen K H, Wang K J, Tsai M L, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm [J]. BMC Bioinformatics, 2014, 15 (1): 49.
[16] Mohamad M S, Omatu S, Deris S, et al. Particle swarm optimization for gene selection in classifying cancer classes [J]. Artificial Life & Robotics, 2009, 14 (1): 16-19.
[17] Chinnaswamy A, Srinivasan R. Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data [M]// Innovations in Bio-Inspired Computing and Applications. [S. l.]: Springer International Publishing, 2016: 1117-1124.
[18] Sheikhpour R, Sarram M A, Sheikhpour R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation

based classifiers in diagnosis of breast cancer [J]. Applied Soft Computing, 2016, 40: 113-131.
[19] Taherkhani M, Safabakhsh R. A novel stability-based adaptive inertia weight for particle swarm optimization [J]. Applied Soft Computing, 2016, 38: 281-295.
[20] Manek A S, Shenoy P D, Mohan M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier [J]. World Wide Web-internet & Web Information Systems, 2017, 20 (2): 135-154.
[21] Shi Y, Eberhart R C. Parameter selection in particle swarm optimization [C]// International Conference on Evolutionary Programming Vii. New York: Springer-Verlag, 1998: 591-600.

Note: Figure translations are in progress. See original paper for figures.