INTERNATIONAL JOURNAL OF ELECTRONICS, MECHANICAL AND MECHATRONICS ENGINEERING
Vol.8 Num.2 - 2018 (1537-1542)

Multiclass Cancer Diagnosis using Firefly Algorithm and K- Nearest Neighbor

Elnaz PASHAEI[1]

Abstract - Among a large number of genes in microarray data sets that characterize the samples, many of them may be irrelevant to the learning tasks. Thus there is a need for reliable methods for gene representation, reduction, and selection, to speed up the processing rate, improve the classification accuracy, and to avoid incomprehensibility due to the high number of genes investigated. Classifying multiclass data sets is usually more difficult than classifying microarray datasets with only two classes. In this paper, we propose a new gene selection and classification strategy based on Firefly Algorithm (FFA) and K- Nearest Neighbor (KNN), suitable for multiclass microarray data sets. This approach is associated with Kruskal-test pre-filtering technique. The FFA is utilized to evolve gene subsets whose fitness is evaluated by a KNN classifier with leave-one-out-cross-validation (LOOCV) schema. The experimental results on three multiclass high-dimensional data sets show that the proposed method simplifies gene signatures effectively and obtains approximately higher classification accuracy compared to the best previously published results.

Keywords: *Gene selection, firefly algorithm, kruskal-test, k- nearest neighbor.*

1.Introduction

The DNA microarray technology simultaneously allows for monitoring and measuring the expression level of a great number of genes in tissue samples. In microarray data sets the number of samples is much smaller than the number of genes. The classification of such data results with the known problem of "curse of dimensionality" and data overfitting. Therefore, for a successful disease diagnosis, it is necessary to select a small number of discriminative genes that are relevant for classification. Gene selection in microarray data analysis, not only increases the classification accuracy, but also decreases the processing time in the clinical setting. Hence, it is quite important to determine a minimum subset of genes for developing a successful disease diagnostic system. There are different methods developed for gene selection in recent years. These methods can be categorized into two main groups as the filter (ranking) and wrapper (gene subset selection) approach. The filter approach assesses each gene individually and ranks the genes from the most relevant to the less relevant using a certain 'filter' criteria. The filter approaches that can be used without restriction in the multiclass case are F-test, Kruskal-test, Random Forest (RF), and boosting. Multiclass generalization to the Wilcoxon rank sum test and the nonparametric pendant to the F-test is known as Kruskal-test. Wrapper approaches evaluate the goodness of each found gene subset by the estimation of the accuracy percentage of the specific classifier.

[1] *Software Engineering, Engineering Faculty, Istanbul Aydin University, Istanbul, Turkey, elnazpashaei@aydin.edu.tr*

The classifier is trained only with the found genes. Wrapper approaches, when compared to the filter approaches, obtain better classification performance, however they are more of a computational cost. Evolutionary algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) [1-3], Ant Colony Optimization (ACO) [4], Binary Black Hole Algorithm (BBHA) [5], and Firefly Algorithm (FFA) [6] are some wrapper based approaches that have been provided and widely applied in bioinformatics. Since these approaches simultaneously evaluate many points on the search space, they can obtain excellent performance in gene expression data analysis. FFA has been used effectively to solve various NP-hard problems such as image processing, shape, and size optimization, set covering problem, manufacturing cell problem and gene selection [6-9]. However, combining FFA with 1NN classifier and applying it as gene selector on gene expression datasets has rarely been investigated by previous researchers. Gene selection and classifier design are known as two crucial factors in determining the performance of gene expression classification problem. In fact, the gene expression classification results depend on selected relevant gene subsets and performance of the classifiers. In classifier design, classification of multiclass (class >2) microarrays are usually more difficult than the classification of microarray datasets with only two classes. The support vector machines (SVMs) [6], nearest Shrunken Centroids Discriminant Analysis (SCDA) [10], Random Forest, and K-nearest neighbor (K-NN) [1] are three prevalent classifiers, which have been found useful in handling classification tasks in the case of the high dimensionality and multiclass data. The K-NN is one of the most popular nonparametric methods that were introduced by Fix and Hodges in 1951. K-NN is invariant to noisy data and not negatively affected when the training data is large. For error estimation on the classifier, the leave-one-out-cross-validation (LOOCV) schema can be utilized. The LOOCV technique is a straightforward and unbiased estimator that is widely used in small sample-sized data sets. In this paper, we are interested in gene selection and the classification of multiclass microarray data. For this purpose, we proposed a hybrid model that uses two techniques: LOOCV Kruskal-test and Firefly Algorithm (FFA) combined with one nearest neighbor (1-NN). First, to cope with the difficulty related to high-dimensional data, we use a Multi-class generalization to the Wilcoxon rank sum test as a pre-filtering step which ranks the genes from the most relevant to the less relevant for gene reduction. From each data set, 1000 tops ranked genes are selected. Second, the FFA combined with a 1NN classifier is used for final gene selection and classification. The gene subsets were measured by the LOOCV mean absolute error of one nearest neighbor (1-NN). Neighbors are calculated using their Euclidean distance.

The proposed approach is experimentally assessed on three long-familiar multiclass microarrays (9-Tumors, 11-Tumors, and Lung-Cancer). Comparisons with eight well-known classifiers and six state-of-the-art demonstrate that our proposed approach yields a minimum number of genes with high prediction performance. The remainder of this paper is organized as follows; we introduce the general scheme of our hybrid model in Section 2. Experimental results and Comparisons are presented in Section 3. Finally, conclusions are given in Section 4.

2. Gene Selection and Classification by FFA/1NN

In this section, we describe the hybrid FFA/1NN algorithm for performing gene selection and classification of multiclass microarray data. The FFA is designed both for identifying optimal gene subsets (solutions) and for final gene selection and classification. The 1NN-based classifier is used to ensure the fitness evaluation of each candidate solution as part of the firefly based wrapper algorithm.

a) The Firefly Algorithm

The Firefly Algorithm (FFA) is a novel nature-inspired algorithm which was presented by Xin-She Yang in 2008 [7] and applied for solving the linear design problem and multimodal optimization problem. The idea of the FFA is to mimic the behavior of flashing lights of fireflies. The FFA was developed by utilizing the following three idealized rules:

- All fireflies are unisex and are attracted to other fireflies regardless of their sex.
- The degree of the attractiveness of a firefly is proportional to its brightness, and thus for any two flashing fireflies, the dimmer firefly is attracted by the brighter one and moves towards it. The fewer distance between two fireflies means more brightness. Fireflies move randomly if there are no brighter fireflies nearby.
- The brightness of a firefly is determined by the value of the objective function.

Based on these three rules the pseudo code of FFA is shown in Figure 1.

Objective function $f(x), x = (x_1, x_2, \dots, x_d)^T$
Generate an initial population of n fireflies $x_i(1, 2, \dots, n)$
Light intensity I_i at x_i is determined by $f(x)_i$
Define a light absorption coefficient $(\gamma) = 0.001$;
Define mutation Coefficient $(alpha)$=0.01;
while $(t < \max generartion)$
for $i = 1{:}n$
for $j = 1{:}n$
if $I_i < I_j$

$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=d}^{1}(x_{i,k} - x_{j,k})^2}$;

$\beta = \beta_0 e^{-\gamma r_{ij}^2}$; $\beta_0 = 0.33$;

$x_i^{t+1} = x_i^t + \beta\left(x_j^t - x_i^t\right) + alpha\ \epsilon_i^t$;

ϵ_i^t is a vector of random numbers drawn from a uniform distribution
end if
Evaluate new solutions and update light intensity
end for j
end for i
Rank the fireflies and find the current global best g∗
end while.

Figure 1. Pseudo code of the firefly algorithm.

b)Fireflies and initial population.

The Fireflies are binary-encoded; each allele (a bit) of the fireflies represents a gene. If an allele is "1" it indicates that this gene is kept in the gene subset and "0" means that the gene is not included in the subset. Thus, each firefly represents a gene subset. The firefly length is equal in the number of genes pre-selected by the Kruskal test pre-processing (i.e. 1000 for each data set). The initial population of the FFA is randomly generated according to a uniform distribution.

c)Objective function.

The fitness of a firefly, i.e. a subset of genes, is evaluated by LOOCV classification mean absolute error of 1NN classifier. In other words, the lower fitness value is gotten; the better gene subset may be obtained.

d)Stopping criterion.

The evolution process ends when a pre-defined number of generations (200) is reached.

3.Evaluation

a)Parameters Settings

Table 1 summarizes three multiclass gene expression data sets that are used for this study. These data sets have thousands of genes (high-dimensional data). They were downloaded from http://www.gems-system.org. All the experimental results reported in this article was acquired using WEKA open source machine learning software and R packages. Firstly, a Kruskal-test was applied for pre-processing in order to pre-select 1000- tops-ranked genes. For performing Kruskal-test, "CMA" package in R [11]was used. The genes were then applied in FFA. Next, the LOOCV mean absolute error of gene subsets that were produced by FFA, was measured by using KNN. Generally, in LOOCV, one sample among all samples is evaluated as testing data while the others are used as training data. This is repeated so that each observation in the sample is used once as the test data. The sizes of population and iterations for all data sets are set to 50 and 200, respectively. These parameters are same for cuckoo search. For FFA, except mutation type that must be set to bit-off, the remaining parameters are set as default.

b)Results and Comparisons

Firstly, in order to accelerate the speed of convergence and alleviate the burden of computation, 1000 top ranked informative genes were selected by Kruskal-test approach. Then to further reduce the number of marker genes and improve the classification accuracy, the FFA/1NN algorithm was applied on these 1000 genes.

Table 2 reports the LOOCV classification accuracy of the five classifiers without using any gene selection approach on 9-Tumor, 11-Tumors, and Lung cancer data sets. The results presented in this table imply that without using any gene selection approach, we cannot be able to capture the patterns that underlie the gene expression profiles.

Table 3 shows the LOOCV classification accuracies of eight different classifiers on 1000 top ranked genes which were obtained by filter-based feature ranking approach (Kruskal-test).We compared the LOOCV classification accuracy of the FFA/1NN algorithm proposed in this paper with the following eight most popular algorithms; Cuckoo search/Naive Bayes, PART, 1NN, Boosted C5.0, Correlation-based Feature Subset selection (CFS)/Multinomial logistic regression, SVM with the polynomial kernel, Random Forest (RF), and SCDA. Experimental results show that our method resulted in higher averages of the classification accuracies on all data sets compared to the eight methods in Table 3.

To carry out our experiments, our FFA/1NN algorithm is run 5 times on each of the 9-Tumor, 11-Tumors, and Lung cancer multi-class microarray data sets (Table 4). Table 5 summarizes our results (Column 2) for these data sets with the results of six state-of-the-art methods from the literature (Columns 3-8). Two criteria are used to compare the results: the classification accuracy (first number) and the number of used genes (the number in parenthesis).

For all the data sets, the averages of the number of the selected genes for our work were smaller than the previous works [1-3, 12, 13]. As it can be observed, for the 9-Tumor data set, we obtained a classification rate of 90.66% using 43.2 genes, which is much better than that reported in [1-3, 12-14]. The study [13] has shown better classification accuracy than our work on Lung cancer data set but with a greater number of genes (99.52% with 6958 genes). Our approach offers the correct classification rate as 98.32% with only 21.8 genes. For 11-Tumors data set, our approach has achieved the highest (averaged) classification accuracy with the minimum number of genes. The same performance is achieved by [13], with a high number of selected genes.

Table 4 shows the detailed results of five independent runs of our FFA/1NN algorithm. As it can be observed, these results are quite stable in all data sets based on the standard deviations. For the 11-Tumors and Lung cancer data sets, each of the five runs obtains a classification rate of 97% and 98 % while for the 9-tumor data set, the best run gives a classification rate of 93.33. Even the worst obtains a classification rate of 88.33.

Experimental results show that our proposed Kruskal-test/FFA/1NN algorithm may select a smaller gene subset with better LOOCV classification accuracy than many other methods in almost all data sets. Therefore, it is more effective for gene subset selection and pattern classification on multiclass data sets.

4.Conclusion

In this paper, a new hybrid algorithm was presented for gene selection and classification of multiclass high dimensional microarray data. The FFA Algorithm employed KNN classifier to intelligently select the most convenient genes that could maximize the classification accuracy while ignoring the redundant and noisy genes. The proposed approach, compared to the existing methods, achieves better classification accuracy with significantly fewer numbers of genes.

References

[1] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry,* vol. 32, pp. 29-38, 2008.
[2] B. Tran, B. Xue, and M. Zhang, "Improved PSO for Feature Selection on High-Dimensional Datasets," *Springer International Publishing Switzerland,* pp. 503–515, 2014.
[3] E. Pashaei, M. Ozen, and N. Aydin, "A Novel Gene Selection Algorithm for cancer identification based on Random Forest and Particle Swarm Optimization," presented at the Proceedings of 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Niagara Falls, Canada, 2015.
[4] Y. Hualong, G. Guochang, L. Haibo, S. Jing, and Z. Jing, "A Modified Ant Colony Optimization Algorithm for Tumor Marker Gene Selection," *Genomics Proteomics Bioinformatics,* vol. 7, pp. 200–208, 2009 Dec.
[5] E. Pashaei and N. Aydin, "Binary black hole algorithm for feature selection and classification on biological data," *Applied Soft Computing,* vol. 56, pp. 94-106, 2017.
[6] A. Srivastava, S. Chakrabarti, S. Das, S. Ghosh, and V. K. Jayaraman, "Hybrid Firefly Based Simultaneous Gene Selection and Cancer Classification Using Support Vector Machines and Random Forests," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, India, 04 December 2012, pp. 485-494.
[7] X. S. Yang, "Firefly algorithm," *Nature-Inspired Metaheuristic Algorithms,* pp. 79-90, 2008.
[8] B. CRAWFORD, R. SOTO, M. OLIVARES-SUAREZ, W. PALMA, F. PAREDES, E. OLGU´IN*, et al.*, "A Binary Coded Firefly Algorithm that Solves the Set Covering Problem," *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY,* vol. 17, pp. 252–264, 2014.

[9] X.-S. Yang and X. He, "Firefly Algorithm: Recent Advances And Applications," *Int. J. Swarm Intelligence,* vol. 1, pp. 36-50, 2013.
[10] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids with applications to DNA microarrays. ," *Statistical Science,* vol. 18, pp. 104-117, 2003.
[11] M. Slawski, M. Daumer, and A. L. Boulesteix, "CMA - A comprehensive Bioconductor package for supervised classification with high dimensional data," *BMC Bioinformatics,* vol. 9, 2008.
[12] A. J. Ferreira and M. r. A. T. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition,* vol. 45, pp. 3048–3060, 2012.
[13] L. Y. Chuang, C. H. Yang, and C. H. Yang, "Tabu search and binary particle swarm optimization for feature selection using microarray data," *J Comput Biol,* vol. 16, pp. 1689–703, 2009.
[14] M. S. Mohamad, S. Omatu, S. Deris, M. Yoshioka, A. Abdullah, and Z. Ibrahim, "An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes," *Algorithms Mol Biol,* vol. 8, pp. 1-11, 2013.