APPLICATION NOTE

Genomics & Informatics
Vol. 9(3) 134-135, September 2011

http://dx.doi.org/10.5808/GI.2011.9.3.134
pISSN 1598-866X eISSN 2234-0742

Implementation of a Particle Swarm Optimization-based Classification Algorithm for Analyzing DNA Chip Data

Xiaoyue Han and Minsoo Lee*

Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-750, Korea

Abstract

DNA chips are used for experiments on genes and provide useful information that could be further analyzed. Using the data extracted from the DNA chips to find useful patterns or information has become a very important issue. In this paper, we explain the application developed for classifying DNA chip data using a classification method based on the Particle Swarm Optimization (PSO) algorithm. Considering that DNA chip data is extremely large and has a fuzzy characteristic, an algorithm that imitates the ecosystem such as the PSO algorithm is suitable to be used for analyzing such data. The application enables researchers to customize the PSO algorithm parameters and see detail results of the classification rules.

Availability: The codes for the developed algorithm may be obtained from the author under consent to intellectual property agreements. The code is a demo version and thus some configuration methods are done through editing configuration files. The results are shown in a text window which will be improved in the future. Contact mlee@ewha.ac.kr if you are interested in the demo version of the codes for possible collaboration.

Keywords: classification, DNA chip data

Introduction

Biological science is being revolutionized by the availability of abundant information regarding complete genome sequences for many different organisms.

DNA chips enable at least thousands to millions of the DNA to be placed on a small spaced area. Biologists can experiment with genes by easily using DNA chips. After the experiments, an analysis process is needed. The data provided by the DNA chips could be represented as a two dimensional matrix, in which one axis represent genes and the other represent samples. The analysis process makes use of the data extracted from the DNA chips and performs various mining operations such as classification to find meaningful patterns in the data.

The classification algorithms for DNA chip data could benefit from using algorithms that imitate the ecosystem. These algorithms are adaptable to large scale problems using probabilistic search mechanisms. Among such algorithms the Particle Swarm Optimization algorithm (Falco *et al*., 2006) has recently emerged as an interesting candidate for dealing with biological data.

In this paper we provide a simple description of the application developed using the PSO algorithm to classify DNA chip data. The parameters for the algorithm can be changed for various experimental data and the details of the resulting classification rules are shown by the application.

Particle Swarm Optimization Algorithm

PSO 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. The PSO algorithm is not only a tool for optimization, but also a tool for representing socio-cognition of human and artificial agents, based on principles of social psychology. Some scientists suggest that knowledge is optimized by social interaction and thinking is not only private but also interpersonal.

The PSO algorithm works as the following (Falco *et al*., 2007). A swarm of particles is created at the beginning. Each particle has a position and velocity value in a multidimensional space. Each position is assigned a value based on an objective function that guides the algorithm. Particles in the swarm exchange good positions among themselves and adjust their own position and velocity based on these good positions. Globally best positions that affect all particles as well as locally best positions which affect a subset of neighboring particles are remembered. As several iterations continue, the globally best position is updated and the final result gives the best particle position.

PSO-based classification system

The PSO algorithm was used to develop the DNA chip

*Corresponding author: E-mail mlee@ehwa.ac.kr
Tel +82-2-3277-3401, Fax +82-2-3277-2306
Accepted 4 June 2011

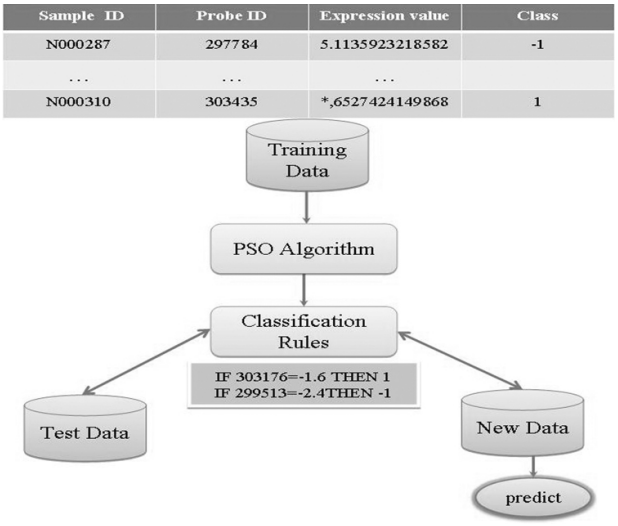

Fig. 1. PSO-based classification system.

data classifying system shown in Fig. 1.

The analysis system can be divided into two phases. The first phase is creating classification rules as explained above and the second phase is applying the classification rules on new data to predict the class. The first phase of discovering classification rules is composed of two steps. The first step is the input pre-processing step where the analysis system gets data from the database and converts it into an appropriate data format so that it can be used in the analysis system. The second step is the training step using the PSO algorithm to create classification rules. An example for the training data is shown in Fig. 1. It contains the sample ID, probe ID, Expression level, and class. The training data and test data are divided with a 10-fold cross validation mechanism. The particles of the PSO algorithm represent rules in the classification system. Each particle includes information such as the probe number, gene expression value, class, local accuracy of rule. The fitness of a particle is defined as the accuracy of the rule it represents. The position of a particle is updated by calculating the difference between the velocity in time t and t+1. When updating the position, a fitness value of each rule is computed. And if the calculated fitness value is greater than the global fitness, then the current fitness replaces the global fitness. After the training process, many rules are produced and the best rules would be chosen as a final set of rules. Then the rule is applied to the test data and from the predicted class results we can compare them with the correct answers provided by the test data set to get the accuracy of the prediction. The second phase is using the rules discovered from the previous step to predict the classes

```
18 the Rule accuracy : 0.000000 %

rule [19] :
IF
298316 = 0.300000, and 299792 = -2.200000, and
301664 = -5.200000, and 301972 = -1.000000, and
300327 = -1.300000, and 302450 = -2.900000, and
303435 = 0.500000, and 303366 = -2.500000, and
300862 = 0.000000, and 302796 = 4.900000
THEN -1, Default 1
19 th Rule accuracy : 75.000000 %


==========<Result of iteration : 500>==========

Accuracy : 100.000000
IF 300142 = -0.300000, and
298604 = -0.900000, and 298143 = -3.900000, and
299330 = 0.800000, and 298276 = 0.500000, and
299931 = -0.800000, and 298655 = 0.700000, and
298923 = -0.100000, and 298704 = 1.400000, and
300922 = 4.500000 THEN -1, DEFAULT 1
```

Fig. 2. Result of PSO-based classification.

for the new data set.

The PSO-based classification system is implemented with C. Fig. 2 shows the results of the classification rules found. The number of iterations can be customized by the user. The DNA chip data used was the AB 1700 mouse chip (1-dye) which was provided from Macrogen Inc. and was composed of 100 different genes and 24 samples. The number of classes is 2 and each class has 12 samples.

Discussion

The PSO-based classification system can be used for DNA chip data analysis. Several customization options are available but more options will be available in the future to customize the fitness functions and particles in some other applications. Some experiments showed that the algorithm usually provides rules with more than 90% accuracy and less than 70% of time to discover the classification rules compared to other traditional approaches such as decision trees. Future work will validate the performance more extensively and provide a more advanced user interface.

Acknowledgements

This work was supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. 2009-0083992).

References

Falco, I.D., Cioppa, A.D., and Tarantino, E. (2006). Evaluation of Particle Swarm Optimization Effectiveness in Classification. *Fuzzy Logic and Applications* 3849, 164-171.

Falco, I.D., Cioppa, A.D., and Tarantino, E. (2007). Facing classification problems with Particle Swarm Optimization. *Appl. Soft. Comput.* 7, 652-658.