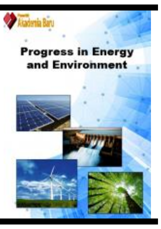
Penerbit Akademia Baru

Progress in Energy and Environment

Journal homepage: http://www.akademiabaru.com/progee.html
ISSN: 2600-7762



Particle Swarm Feature Selection for Microarray Leukemia Classification

Research Article

Win Son Ng[1], Siew Chin Neoh[*,1],Kyaw Kyaw Htike[1], Shir Li Wang[2]

[1] Faculty of Engineering, Technology, and Built Environment, UCSI University, Kuala Lumpur, Malaysia
[2] Sultan Idris Education University, Tanjong Malim, Perak Darul Ridzuan, 35900 Malaysia

ARTICLE INFO

Article history:
Received 12 September 2017
Received in revised form 05 October 2017
Accepted 05 October 2017
Available online 26 October 2017

Keywords:
Lymphoblastic Leukemia, Feature Selection, Particle Swarm Optimization, Microarray

ABSTRACT

In the recent years, DNA microarray has been widely used to investigate genes that cause genetic diseases. Since information from DNA microarray could reveal some interesting relationships between genes and diseases, it has been employed by a number of researchers to classify Acute Lymphoblastic Leukemia (ALL) and Acute Myelogenous Leukemia (AML). As microarray gene expression involves high dimensional features, feature reduction or feature selection is required to ensure efficient classification of ALL and AML. This paper proposes a multi-population particle swarm optimization (MPSO) feature selection approach to identify the most significant subsets of genes for classification of ALL and AML. In this research, MPSO is used to increase the search diversity of conventional particle swarm optimization (PSO). It is combined with the Support Vector Machine (SVM) classifier to form a wrapper feature selection model that can capture the interactions between the classifier and the features. The proposed model is evaluated using 10-fold cross validation. Results showed that MPSO gives a more consistent classification performance than the conventional PSO in ALL and AML classification.

Copyright © 2017 PENERBIT AKADEMIA BARU - All rights reserved

1. Introduction

Classification of acute lymphoblastic leukemia (ALL) and acute mylogenous leukemia (AML) is a critical and important factor for successful leukemia treatment. The classification methods of ALL and AML have gone through a few evolutions throughout the decades. These methods include making observations on tumors and blood testing, and performing image processing on the image captured under the view of microscope. However, there is a shortage of a systematic method and approach in discovering the leukemia class even though the classification of cancer has achieved constant improvements over the past years [1].

Nowadays, ALL and AML can be classified by exploring and analyzing the DNA and RNA data from human body. DNA microarray is commonly used to provide gene expression. The gene

[*] *E-mail address: u_jane80@yahoo.co.uk*

expression often contains high dimensional of data with a small number of samples. As a result, computational biology with feature selection is often used to analyze the collected gene expression in order to assist therapists and scientists for making more accurate judgments. According to [2], one of the major challenges of feature selection is the involvement of high number of features with a small samples size [2]. This challenge has attracted research attention towards the development of evolutionary machine learning algorithms to stochastically identify suitable features for classification purposes.

This research adopts the concept outlined by PSO to select relevant gene features that could distinguish ALL and AML through SVM classification. Overall, the paper is organized in the following way. Section 2 reviews the literature related to feature selection. Section 3 presents the proposed MPSO feature selection for ALL and AML classification using SVM classifier whereas Section 4 discusses the results of MPSO as compared to conventional PSO. Lastly, Section 5 concludes the findings obtained from the proposed method and identifies the future research direction.

2. Evolutionary Algorithm in Feature Selection

Feature selections are commonly used in pattern recognition and data mining. In general, feature selection models can be categorized into three major categories [2-6]: (i) Filter, (ii) Wrapper, (iii) Embedded. Filter model conducts feature selection process based on the general characteristics of training data without including interaction with classifier whereas wrapper model optimizes a classifier or predictor as part of the feature selection process [6-8]. On the other hand, embedded model performs feature seletion by adding additional constraints into optimization of the predictive algorithm [9]. The computational cost is normally lowest for the filter model and the highest for wrapper model. Though computationally intensive, wrapper model permits black box machine learning and are able to capture feature dependencies [10].

Due to the attractive properties of evolutionary algorithms (EAs) which could improve feature selection through recursive optimization of classifier accuracy, EAs such as Genetic Algorithm (GA), PSO, Artificial Bee Colony (ABC) algorithm and etc are commonly applied as wrapper-based feature selection algorithms. For instance, Chen et al. [11] applied a coarse-grained parallel GA to simultaneously obtain the optimized feature subset and parameters for SVM whereas Aalaei et al. [12] proposed GA as a wrapper feature selection approach for breast cancer diagnosis. Zhang et al. [13] proposed a multi-label feature selection algorithm using an improved multi-objective PSO while Mistry et al. [14] employed a micro-GA embedded PSO feature selection for facial emotion recognition. Besides GA and PSO, Subanya and Rajalaxmi [15] applied ABC to optimize feature subset for cardiovascular disease classification.

This research focuses on PSO as the wrapper-based feature selection technique to classify ALL and AML. As conventional PSO tends to trap in a local optimum when large search space is encountered, this paper proposed MPSO to increase the search diversity and inject more randomness into the feature selection model so as to distinguish AML and ALL more efficiently.

3. Research Methodology

This paper proposes a wrapper feature selection model to distinguish ALL and AML. The basic operation of the proposed wrapper model is shown in **Fig. 1**. In the model, MPSO is developed as the feature selection algorithm whereas SVM is used as the classifier to identify ALL and AML. There is a recursive update on the relevant gene features to be used for better SVM classification. From iteration to iteration, the classification accuracies are fed to MPSO to update gene subset (feature selection). Subsequently, the updated feature subsets are again used for classification whereby the interaction process is repeated until the maximum number of iterations is met.

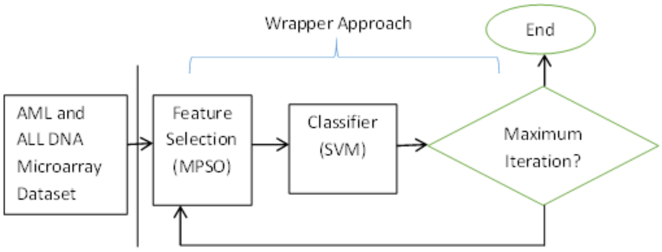


Fig. 1. The basic operation flow of MPSO wrapper feature selection model

3.1 *Multi-population Particle Swarm Optimization (MPSO)*

The idea of PSO is inspired by the food hunting behavior of a group of birds in a specific area. In PSO, each bird (also called as particle) in the swarm is given a fitness score based on the amount of food found. Birds with more food will have higher fitness score as compared to the other. Based on the score, birds change their flying direction and move towards the direction where better food source is expected. From time to time, birds or particles update its location, flying speed and food score in order to allow other birds to learn from them and further explore for better food source.

In this research, the gene expression dataset for ALL and AML classification is obtained from the Cancer Program Legacy Publication Resources of BRPAD Institute [1]. 72 samples with 7219 genes are used in this study. With the analogies of particle swarm mentioned above, several gene subsets are randomly initialized as particles in the swarm. Each of these gene subsets is fed to the SVM classifier to determine the classification accuracy which will indicate the fitness score of the particle.

In the evolution of PSO, particle with highest fitness value in the swarm is called *gbest* whereas the best fitness value a particle has ever encountered is called *pbest*. Both *gbest* and *pbest* are used to guide the search of particles towards better food source. In other words, better food source means better fitness score which indicates better classification accuracy with smaller gene subset. Equation (1) shows the velocity update of particle based on *gbest* and *pbest* whereby r_1 and r_2 are the random numbers generated in between [0, 1] whereas c_1 and c_2 are the control variables that manipulate weightage of *pbest* and *gbest* respectively. In the equation, *i* refers to particle i^{th}, *d* refers to the d^{th} feature or gene while *v* refers to the velocity.

$$v_{id}^{new}=w\times v_{id}^{old}+c_1\times r_1\times\left(pbest_{id}-x_{id}^{old}\right)+c_2\times r_2\times\left(gbest_d-x_{id}^{old}\right) \quad (1)$$

After updating the velocity, the particle's position, *pos*, is updated using (2) where the new position for the particle is obtained by adding the old position with velocity.

$$pos_{id}^{new}=pos_{id}^{old}+v_{id}$$
(2)

All the particles move to the new location according to the newly generated velocity. Then, the new updated particles are again evaluated for fitness to identify the next *pbest* and *gbest* for upcoming velocity updates. Equation (3) shows the fitness evaluation function where f_i refers to the fitness of particle *i* and *Acc* refers to the classification accuracy.

$$f_i=Acc \quad (3)$$

Using (3), each particle is updated with latest fitness score and learns from other particles in the swarm to update its velocity and position using (1) and (2). After updating the position, the particle is again evaluated for fitness. This process is repeated until the maximum iteration is reached. **Fig. 2** shows the flow chart of general PSO.

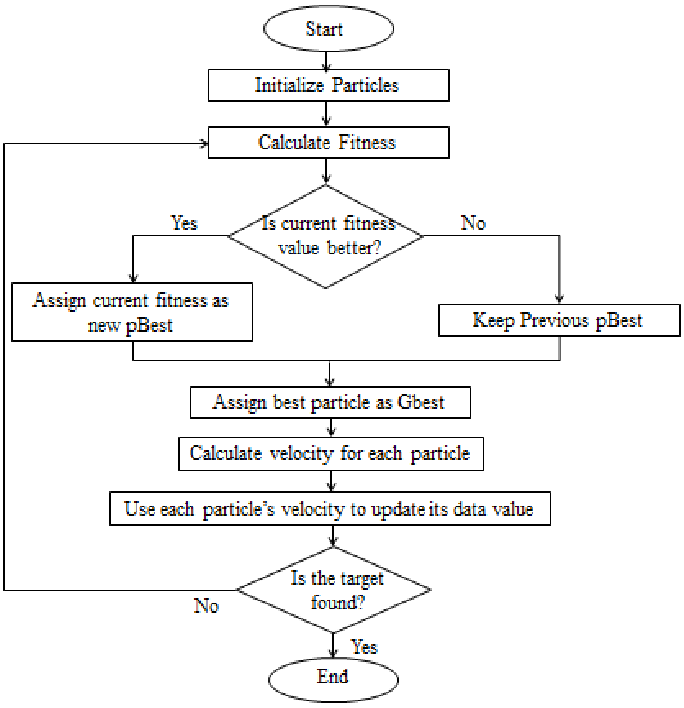

Start
Initialize Particles
Calculate Fitness
Is current fitness value better?
Yes
No
Assign current fitness as new pBest
Keep Previous pBest
Assign best particle as Gbest
Calculate velocity for each particle
Use each particle's velocity to update its data value
Is the target found?
No
Yes
End

Fig. 2. Flow Chart of General PSO Algorithm

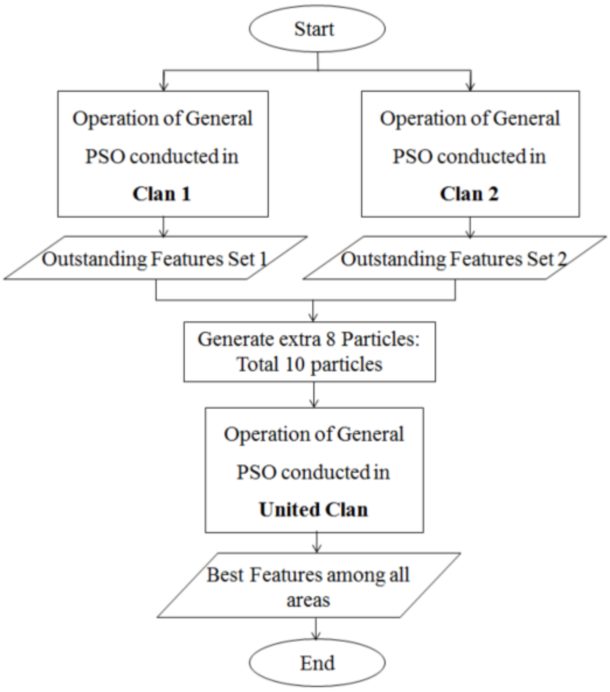

Start
Operation of General PSO conducted in Clan 1
Operation of General PSO conducted in Clan 2
Outstanding Features Set 1
Outstanding Features Set 2
Generate extra 8 Particles: Total 10 particles
Operation of General PSO conducted in United Clan
Best Features among all areas
End

Fig. 3. Flow Chart of Multi-population PSO

The major weakness of the general PSO is the high tendency to trap in the local optimum when large search space is involved with small population size of particles. The main reason of this issue is the incapability to explore for more diverse solutions. In order to increase the chances to explore for more diverse solutions, MPSO is proposed in this research to diverse the search of general PSO. Instead of generating only a single swarm where all particles follow the same leader, multiple swarms are generated to move particles in several different directions since each swarm will have a global best (*gbest*) swarm leader.

In this paper, we proposed to use two swarms, called clan 1 and clan 2 respectively for the search of best feature subsets. In order to enhance exploration, the features which are selected in clan 1 should not be selected for clan 2. For each clan, the particles will update positions based on the best leader in the clan and their personal best experience. Therefore, it is possible that particles in different clan will fly in different directions according to the conditions in their own clan for a number of iterations.

After a maximum number of iterations, each clan contributes its best particle to a united clan. Besides the two particles contributed by clan 1 and clan 2, united clan random generate another 8 particles in its population to infuse more randomness for better exploration, aiming to avoid the local optimum trap. The swarm in united clan is then continued with the general velocity and position updates until it reached the maximum iteration. After termination of the search, the position of the particle with the best fitness is taken as the best gene subset for ALL and AML classification. **Fig. 3** shows the flow chart of Multi-population PSO. The operation of MPSO can be considered as a tournament where the best candidate from each clan is inserted to a united clan or finale to search for the best of the best.

3.2 *Support Vector Machine (SVM)*

SVM is a classifier that is able to perform discriminative classification based on hyperplane. It is popular in determining the dichotomy classification problems [16]. SVM classifier differentiates the data based on two processes: Training and Testing.

Training process: The SVM classifier begins with the plotting of the selected data from PSO and mapped into a graph according to the features. SVM classifier determined the best hyperplane that used to classify ALL and AML according to the actual diagnosed results of the patients. This hyperplane is known as the SVM structure.

Testing process: The best structure determined from training session is used to test the unseen data in the test stage. The main purpose of testing is to validate whether the similar SVM structure obtained in the training stage is applicable to unseen data. The test accuracy is determined by comparing the prediction from the Testing process with the actual diagnosed results of the patients.

In this research, the SVM classification accuracy is used to indicate the particle’s fitness. As each particle represents a feature subset, data samples will be sorted based on the selected features. After sorting the data accordingly, *10-fold* cross validation is used to separate the samples into ten portions where nine portions of the samples will be used for training and one portion for testing. The validation is repeated for 10 times to obtain the average classification accuracy which in turns indicate the particle’s fitness.

4. Results and Discussion

The developed MPSO is compared with general PSO to evaluate its performance in selecting discriminative features to classify ALL and AML. **Table 1** and **Table 2** show the results of MPSO and general PSO respectively. From **Table 1**, it is observed that MPSO gives an average classification accuracy of 80.66% with standard deviation of 2.9617%. On the other hand, for general PSO, the average classification accuracy is 78.95% and standard deviation is 6.3338% (**Table 2**).

Table 1 SVM Classification Accuracy Based on MPSO Feature Selection

Trial Number	Classification Accuracy (%/100)	Number of Features
1	0.8024	4
2	0.769	4
3	0.8458	1
4	0.7899	3
5	0.7607	5
6	0.8339	4
7	0.8393	1
8	0.794	2
9	0.8012	3
10	0.8298	4
Average	**0.8066**	3.1
Standard Deviation	**0.029617**	
Median	**0.8018**	

Table 2 SVM Classification Accuracy Based on General PSO Feature Selection

Number of Trial	Classification Accuracy (%/100)	Number of Feature
1	0.7631	4
2	0.7262	5
3	0.9012	3
4	0.7946	3
5	0.8482	2
6	0.7036	4
7	0.7583	1
8	0.7345	5
9	0.8458	3
10	0.8196	2
Average	**0.78951**	3.2
Standard Deviation	**0.063338**	
Median	**0.77885**	

Table 3 Fisher's F-Test for Variances Comparison between MPSO and General PSO

Ratio	0.219
F (Observed value)	0.219
F (Critical value)	0.315
DF1	9
DF2	9
p-value (one-tailed)	0.017
alpha	0.05

Though the mean for both models seems to be close to each other, it is observed that the standard deviation of classification accuracy from general PSO doubled the proposed MPSO. In order to convince the observation, Fisher's F-Test (**Table 3**) is used for two sample comparison of variances based on 95% confidence interval. Two hypotheses are created for the test: (1) Null hypothesis, H0 and (2) Alternative hypothesis, H1.

H0: Ratio of Variance of MPSO to general PSO is equal to 1.
H1: Ratio of Variance of MPSO to general PSO <1.

From **Table 3**, the computed p-value is lower than the significance level, alpha=0.05. Therefore, null hypothesis is rejected and alternative hypothesis is accepted. The risk to reject null hypothesis while it is true is 1.7%. As a result, the test shows 95% confidence that the variance of MPSO is less than general PSO. Thus, MPSO ensure more consistent exploration as compared to general PSO which might trap in local optimum and give results with larger variance.

5. Conclusions

In this paper, PSO is combined with the SVM classifier to distinguish ALL and AML using DNA microarrays data set provided by BROAD Institute. In order to ensure more diversity in the search of general PSO, MPSO is developed with the aim to balance global and local search for better and more consistent classification accuracy. In MPSO, two swarms are generated with different selection of features to enhance diversity. Each swarm evolves based on its leader and converged. After a maximum number of iteration, the best solution from each swarm is united together with another set of randomly generated particles to form a united clan. The search again evolves to obtain the best combination of genes for ALL and AML classification. MPSO is compared with general PSO in terms of classification accuracy and variance of results. From the evaluation, MPSO has shown higher average accuracy with less result variance as compared to general PSO. Overall MPSO provide more consistent search than general PSO due to diverged exploration from multiple swarms. In future, several swarm architectures and parameters will be investigated to improve feature selection for better classification accuracy.

Acknowledgement

Authors would like to express our gratitude to UCSI University for the supports and assistance given.

References

[1] Golub, T.R., Donna, K.S., Pablo, T., Christine, H., Michelle, G., Jill, P.M., Hilary, C., Loh, M.L., Downing, J.R., Caligiuri, M.A. and Bloomfield, C.D. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286, no. 5439 (1999): 531-537.
[2] Bolón-Canedo, V., Noelia, S.M., and Amparo, A.B. "A review of feature selection methods on synthetic data." *Knowledge and information systems* 34 (3) (2013): 483-519.
[3] Singh, R.K. and Sivabalakrishnan, M. "Feature selection of gene expression data for cancer classification: A review." *Procedia Computer Science* 50 (2015): 52-57.
[4] Das, S.F., "Wrappers and a Boosting-Based Hybrid for Feature Selection." In *Proceedings of 18th International Conf. on Machine Learning*, pp. 74-81, 2001.
[5] Ng, A.Y. "On Feature Selection: Learning with Exponentially many Irrelevant Features as Training Examples." In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 404-412, 1998.
[6] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, pp. 856-863, 2003.
[7] Hira, Z.M., and Duncan, F.G. "A review of feature selection and feature extraction methods applied on microarray data." *Advances in Bioinformatics* (2015).
[8] Lal, T.N., Chapelle, O., Weston, J. and Elisseeff, A. "*Embedded methods*." Feature extraction. Springer Berlin Heidelberg, pp. 137-165, 2006.
[9] Iglesia, B.D.L. "Evolutionary computation for feature selection in classification problems." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3 (6) (2013): 381-407.
[10] Chen, Z., Lin, T., Tang, N. and Xia, X. "A parallel genetic algorithm based feature selection and parameter optimization for support vector machine." *Scientific Programming* (2016).

[11] Aalaei, S., Shahraki, H., Rowhanimanesh, A., and Eslami, S. “Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets.” *Iranian Journal of Basic Medical Sciences* 19 (5) (2016): 476-482.

[12] Zhang, Y., Gong, D., Sun, X. and Guo, Y. “A PSO-based multi-objective multi-label feature selection method in classification.” *Scientific Reports* 7 (2017) Article number: 376

[13] Mistry, K., Li, Z. and Neoh, S.W. “A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition.” *IEEE Transactions on cybernetics* PP (99) (2016): 1-14.

[14] Subanya, B. and Rajalaxmi, R.R. “Feature selection using Artificial Bee Colony for cardiovascular disease classification.” In *IEEE International Conference on Electronics and Communication Systems* (2014).

[15] Rakkeitwinai, S., Lursinsap, C., Aporntewan, C. and Mutirangura, A. “New feature selection for gene expression classification based on degree of class overlap in principal dimensions.” *Computers in biology and medicine* 64 (2015): 292-298.