Review on the Usage of Swarm Intelligence in Gene Expression Data

Nurhawani Ahmad Zamri, Bhuvaneswari Thangavel, Nor Azlina Ab Aziz, and Nor Hidayati Abdul Aziz

Abstract

This paper presents a review of the recent usage of swarm intelligence for optimizing feature selection in microarray data focusing on its application for cancer detection and classification. The feature selection technique is used in the analysis of microarray so that only useful data is trained for further analysis and prediction. The process of feature selection would affect the effectiveness of the classification. This is due to the enormous quantity of genes being expressed at the same time. An optimized feature selection would ensure a high accuracy of classification. Swarm intelligence has been effective in solving feature selection and classification problems. This paper also gives overview on the sources of microarray data which are used in the literature.

Keywords

Swarm intelligence • Microarray • Feature selection • Classification

1 Introduction

Swarm intelligence (SI) is a computational model which is commonly inspired by the social behavior seen in nature. It is based on the collective behaviors resulting from the local interactions of individual agents with each other and their environment. The agents act in a coordinated way even in the absence of external controller or coordinator.

SI can be applied in a variety of fields including resource management such as nurse scheduling problem [1], information technology like data clustering [2] and engineering such as path tracking of autonomous mobile robot [3]. Application of SI in the field of bioinformatics as optimizers for feature selection algorithms is the focus of this paper. The usage of SI in this field had shown good results [4].

Feature selection in microarray data is a dimensionality reduction problem. Feature selection plays a crucial role in ensuring efficient analysis of the high dimensionality microarray data. Numerous studies have shown that most genes measured in DNA microarray experiment are not contributing to the accuracy of classification [5]. Therefore selection of genes that highly expressed the disease is essential.

Generally, microarray data is represented in a matrix where rows corresponds to genes and columns corresponds to different samples such as experiment conditions or tissues. Therefore, a cell represents an expression of a gene for a sample. Under various conditions, the transcription levels of genes in an organism is being measured and gene expression is built up. The expression level of a particular gene is expressed as number. The general procedure of DNA microarray analysis can be divided into two main steps as illustrated in Fig. 1. The first step is the selection of the most relevant features. There are numerous feature selection methods that can be applied to the microarray data which can be categorized into filter, wrapper, embedded and hybrid. Alternatively, SI can be applied as optimizers to the feature selection techniques to further improve the selection process.

N. Ahmad Zamri · B. Thangavel (✉) · N. A. Ab Aziz · N. H. Abdul Aziz
Faculty of Engineering & Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
e-mail: t.bhuvaneswari@mmu.edu.my

F. Ibrahim et al. (eds.), *2nd International Conference for Innovation in Biomedical Engineering and Life Sciences*, IFMBE Proceedings 67, https://doi.org/10.1007/978-981-10-7554-4_27

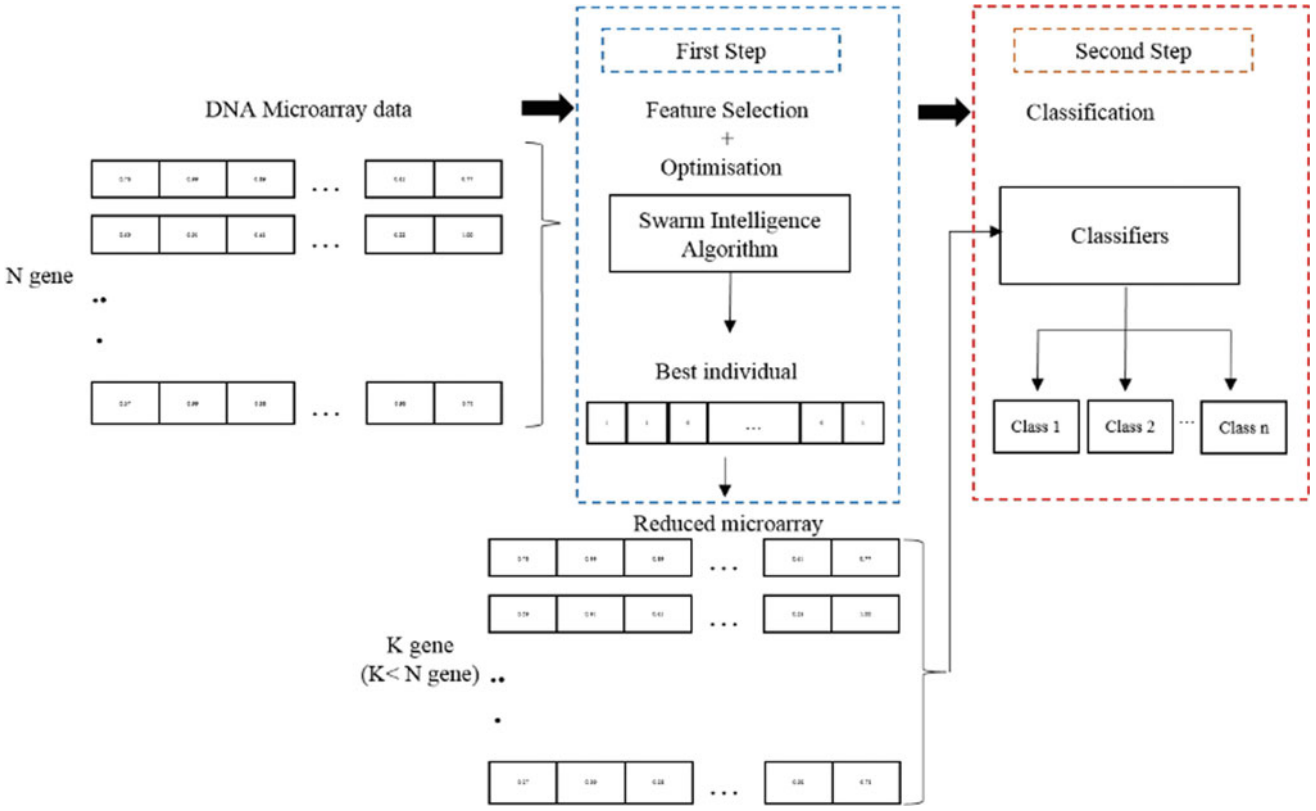

Fig. 1 General procedure of microarray analysis

The best individuals are referred to the genes that best describe the disease. In filter techniques, the interaction with classifier is ignored. A feature relevance score is calculated and low scoring features are eliminated. Then, the feature subset with high scores will be the input to the classifier afterwards. On the other hand, in wrapper techniques, once the best set of genes is obtained, it will be used as training samples for the classifier at the second step to achieve the best classification results. Whereas for embedded techniques, the searching for an optimal subset of features is included with the classification model. A hybrid technique usually combined the feature selection techniques for example a filter and wrapper approach. The evaluation is normally measured in terms of accuracy, specificity and sensitivity.

The structure of the paper is explained as follows: General algorithm of SI and SI optimizers that are frequently used in microarray's feature selection process, such as Particle Swarm Optimization and Artificial Bee Colony Optimization is described in Sect. 2. Section 3 describes the related works done by research community on feature selection using swarm intelligence approach. Section 4 discusses the data source used by the works reviewed in the paper. Finally, Sect. 5 concludes the paper.

2 SI Optimizers

The review presented in this paper is based on eleven works published between 2010 and 2016. It is observed that PSO, ACO and ABC are the popular choice of swarm intelligence algorithms applied to microarray data. Figure 2 shows summary of works reviewed according to the swarm intelligence algorithm used. Based on the literature, PSO is the most popular option with six works and this is followed by

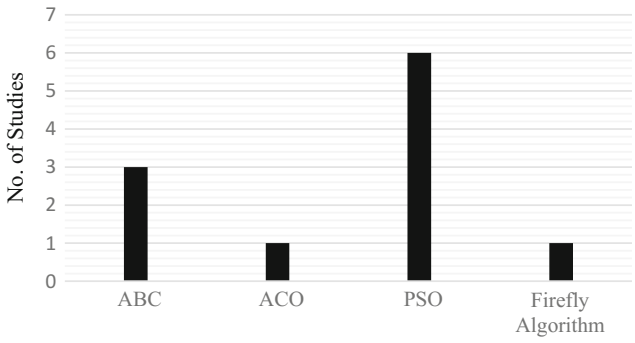

Fig. 2 Recent trends of optimization algorithm used in DNA microarray data

three ABC based research, and one each for ACO and Firefly Algorithm.

The SI algorithms follow similar framework (Fig. 3), where the search for optimal solution is conducted by swarm of agents. Each of the agents holds a candidate solution. The search starts with random initialization of the agents according to the problem. This is followed by evaluation of the quality of the candidate solutions proposed by the agents. The third step is the generation of new candidate solutions, which differs from one algorithm to another. The generation of new candidate solution follows the algorithms' sources of inspiration. For example, in PSO this step mimics the fish schooling and bird flocking behaviour, while the bees swarming behavior is mimicked in ABC, ants foraging is mimicked in ACO and fireflies' flash intensity attraction is mimicked in firefly algorithm. The candidate solutions' quality evaluation and generation of new candidate solutions are repeated until stopping condition is met.

2.1 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) was originally proposed by James Kennedy and Russell Eberhart in 1995 [6]. PSO is a population-based stochastic optimization technique inspired by the social practices watched over animals for example bird flocking and fish schooling. The term particle was used to notate the PSO's search agents [7]. It also refers to population members which are mass-less and volume-less or with an arbitrarily small mass or volume.

A particle, *i,* in the swarm searches for the solution in the high-dimensional problem space using four vectors, its current position, x^i, which represents a solution, the particle's best found position, p^i, the best position found by its neighbourhood from the start of the search, p^g, and lastly its velocity, v^i. The velocity and position at kth iteration is updated using the equations below:

$$v_k^i = v_{k-1}^i + {}^{c_1} r_{1,k}\left(p_k^i - x_{k-1}^i\right) + {}^{c_2} r_{2,k}\left(p_k^g - x_{k-1}^i\right) \quad (1)$$

$$x_k^i = x_{k-1}^i + v_k^i \quad (2)$$

In Eq. (1), c_1 and c_2 represent cognitive and social parameters, while r_1 and r_2 represent random numbers between 0 and 1. PSO has been successfully applied in many areas other than bioinformatics including industrial applications [8], electromagnetic [9] and power systems [10].

```
1: Initialization
2: Evaluate candidate solutions' quality
3: Generate new candidate solutions
4: Check for stopping condition
```

Fig. 3 Pseudo code of SI algorithm

2.2 Artificial Bee Colony (ABC)

Karaboga and Basturk proposed Artificial Bee Colony (ABC) in 2007 [11]. ABC is influenced by the collective behaviour of honey bees to find food sources around the hive. The colony of artificial bees in ABC algorithm contains three groups of bees which are the employed bees, onlookers and scouts.

Employed bees bring loads of nectar to the hive. The employed bees share the information with onlooker bees which stay in the hive. The scout bees search for new food source in the surrounding area of hive. Once an onlooker bee and scout bee select a food source they become employed. The food source is chosen based on a probability value of the following equation:

$$p_i = \frac{fit_i}{\sum_{n=1}^{N} fit_n} \quad (3)$$

where N represents number of food sources and fit_i represents the fitness value of the solution. On the other hand, an employed bee is turned into a scout or onlooker bee when their food sources become empty. A new candidate of food source is calculated by using:

$$x_i^j = x_{\min}^j + rand(0,1)^* \left(x_{\max}^j - x_{\min}^j\right) \quad (4)$$

where x_i^j represents the position of the food source and $x_{\max}^j$ and $x_{\min}^j$ represents the lower bound and upper bound of the j dimension respectively. The possible solution to a problem under consideration is represented by the food source in ABC algorithm in an iterative search process. In ABC algorithm, exploration is done by scout bees whereas the employed and onlooker bees are responsible for exploitation.

3 Feature Selection Using SI

In this section, the related works in feature selection is reviewed. The related works reviewed here are grouped according to the optimization algorithms used.

3.1 PSO Based

PSO has been used extensively within these five years as referred to Fig. 2. This might due to the easy implementation of PSO as it requires few parameters to adjust. The search is decided by the speed of the particle. During the development

of several generations, only the most optimist particle can transmit information onto the other particles [12]. Thus, with this optimization ability, problem can be completed easily.

Particle Swarm Optimization (PSO) was used in [13] for producing an optimized feature subset. The proposed method is tested on four cancer datasets; leukemia, colon, Diffuse Large B-Cell Lymphoma (DLBCL) and breast cancer. The use of PSO during feature selection improves the classification accuracy compared to without using PSO with classification accuracy more than 80%. In [13], Support Vector Machine (SVM) and K-Nearest Neighbourhood (KNN) were used as the classifier.

PSO was also employed in [14]. In [14], PSO and adaptive KNN technique are combined and used for classification of cancer subgroups. Even with least amount of genes in the subset, the classification accuracy still can be improved due to the fitness function of PSO that has been formulated in this work. This is achieved by selecting the optimized value of neighbours, k for KNN classifier. A proper value of k would help the formation of suitable numbers of neighbourhood to be explored. Thus, dataset can be more accurately classified. Less number of genes are appropriately formed by the proposed method from a dimensionally large microarray gene expression data. Results shows that it achieved high classification accuracy on blind test samples.

Hamming distance is applied in [15] compared to normal Euclidian distance measurement. Two different objects may look-alike in an enormous feature space, therefore, the work claimed that Euclidian distance might not be suitable for high dimensional data. The important features subset was selected by the velocity update of particle in binary PSO framework (HDBPSO). In this work, the velocity is updated by the adapted Hamming distance as a proximity measure. The results are evaluated based on validity indices and classification accuracies. Binary data were used in the proposed method for the comparison with GA and multi-objective GA. The results shows HDBPSO performs excellently as compared to others.

In [16], a multi-objective Binary Particle Swarm Optimization (BPSO) algorithms is introduced for feature selection from high dimensional gene expression data. Dimensionality reduction took place at the first stage by which normalization, discretization and conversion of data to binary distinction table occurred. Then, the selection of feature subset occurred at the second stage done by BPSO. Standard classifiers SVM and KNN were used to validate the selected feature subset in terms of classification accuracy.

An integration of PSO searching algorithm and C4.5 decision tree classifiers called PSODT was introduced in [17]. The important genes were identified using PSO algorithm and the fitness of the selection is evaluated using C4.5. The results shows high classification accuracy with PSODT compared to stand alone C4.5 classifier.

A Hybrid Particle Genetic Swarm Optimization (PGSO) is proposed in [18] to optimize the selected features to efficiently classify either normal or early or different stages of ovarian cancer. The basic framework of this proposed work is PSO. However, to improve the PSO performance, Genetic Algorithm (GA) is used as a local optimizer at each iteration. Operation such as selection, crossover and mutation are applied to the initial created particles. The results show an accuracy of 98% with two datasets when using multiclass SVM classifier compared to ANN, 95% and Naïve Bayes, 93%.

3.2 ABC Based

Artificial Bee Colony algorithm was chosen for gene selection problems in several works. Its popularity is also contributed by its few parameters compared to other optimization algorithms. Both global and local search can be conducted by ABC. Hence, the probability of finding the optimal is significantly increased. In [19], ABC algorithm was used to find a subset of genes which is then used to train different Artificial Neural Network (ANN); Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF) and SVM.

A hybrid algorithm called Genetic Bee Colony (GBC) algorithm is introduced in [20]. GBC integrates the ABC algorithm and Genetic Algorithm (GA). In this proposed method, the replacement rate is increased to improve the movement speed. This is done by increasing the number of scout bee from one to two. A mutation operator from GA is also adopted to upgrade the exploitation process of the algorithm. This GBC selection approach is combined with SVM as the classifier. The proposed method shows efficient results in terms of classification accuracy.

Minimum Redundancy Maximum Relevance (mRMR) approach combined with ABC algorithm (mRMR-ABC) is introduced in [21]. mRMR which can be used for continuous and discrete datasets is implemented to measure the relevancy and redundancy of features. This technique can determine the promising features of the gene and facilitate the classifier to be trained accurately. The classification accuracy of SVM classifier will determine the fitness value of the ABC algorithm implemented in this work. The algorithm changes its solution to the new solution when the new fitness value is better than the old fitness values, otherwise it stays in its solution.

Table 1 Data source and accuracy using SI algorithms

Technique	Classifier	Dataset/accuracy (%)	Type of cancer class (binary/multiclass)	Source of data
PSO [13]	SVM	DLBCL (100)	Multiclass	[26]
		Leukemia (100)	Multiclass	
		Breast cancer (100)	Binary	
		Colon (97.50)	Binary	
PSO [17]	Decision tree	11_Tumors (97.82)	Multiclass	[27]
		14_Tumors (74.60)	Multiclass	
		9_Tumors (74.00)	Multiclass	
		Brain_Tumor1 (57.03)	Multiclass	
		Brain_Tumor2 (86.06)	Multiclass	
		Leukemia2 (100)	Multiclass	
		SRBCT (92.94)	Multiclass	
		DLBCL (92.55)	Multiclass	
		5Prostate_Tumor (94.31)	Binary	
		Lung cancer (100)	Binary	
PSO [14]	KNN	SRBCT (100)	Multiclass	[28]
		MLL (100)	Multiclass	
		ALL_AML (97.0588)	Binary	
PSO and GA [18]	SVM	Ovarian cancer (98)	Multiclass	[29]
Multi-objective PSO [16]	KNN, Bayes family function based classifiers and tree based classifiers	Leukemia (>90)	Multiclass	[30]
		Lymphoma (>90)	Multiclass	[31]
		Colon (>90)	Multiclass	[32]
Binary PSO with haming distance [15]	KNN	Leukemia (100)	Multiclass	[30]
		Lymphoma (100)	Multiclass	[31]
		Colon (100)	Multiclass	[32]
ABC and GA [20]	SVM	Lymphoma (98.48)	Multiclass	[27]
		Leukemia2 (95.83)	Multiclass	
		SRBCT (96.38)	Multiclass	
		Leukemia1 (96.43)	Binary	
		Colon (95.64)	Binary	
		Lung (99.50)	Binary	
ABC and mRMR [21]	SVM	Lymphoma (96.96)	Multiclass	[27]
		Leukemia2 (96.12)	Multiclass	
		SRBCT (96.30)	Multiclass	
		Leukemia1 (95.83)	Binary	
		Colon (94.17)	Binary	
		Lung (98.95)	Binary	

(continued)

3.3 ACO Based

ACO is originally developed to solve discrete combinatorial optimization problems such as the travelling salesman problem (TSP). Many of bioinformatics problems like the sequence alignment, gene mapping and feature selection of the gene in microarray data are similar to TSP. This makes ACO suitable for optimization of bioinformatics problems.

Table 1 (continued)

Technique	Classifier	Dataset/accuracy (%)	Type of cancer class (binary/multiclass)	Source of data
ABC [19]	ANN	ALL_AML (91.9)	Binary	[33]
		DLBCL-NIH (59)	Binary	[34]
		Breast cancer (62.1)	Binary	[35, 36]
		Prostate tumor (86.9)	Binary	[37]
ACO and ABC [22]	Fuzzy	ALL_AML (98.7)	Binary	[33]
		Colon (99.5)	Binary	[38]
		Lymphoma (98.5)	Binary	[39]
Firefly Algorithm [24]	NA	National Cancer Institute (NCI) dataset—leukemia, melanoma, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers (NA)	Multiclass	[40]
		Lung dataset (NA)	Binary	

A hybrid Ant Bee Algorithm (ABA) which combines Ant Colony Optimization (ACO) with Artificial Bee Colony (ABC) algorithm in fuzzy expert system for encoding the solution variables using a modified form of representation was introduced in [22]. The optimal rule set of the combinatorial optimization is formed by the implementation of ACO in proposed work. The representation of the membership function as continuous number is done by ABC. The capability of the method is determined using several gene expression data sets which include rheumatoid arthritis versus controls (RAC) and rheumatoid arthritis versus osteoarthritis (RAOA), leukemia, type 2 diabetes, lymphoma and colon cancer. Receiver Operating Characteristic (ROC) analysis have been done to every datasets in the proposed work. The value of area under ROC curve is used to compare the performance of proposed ABA with other algorithm. ABA is reported to have the best value when compared to BCGA, RCGA, PSO and GA.

3.4 Firefly Algorithm Based

Firefly algorithm has the advantage of high convergence rate and robustness [23]. In [24], a multi objective firefly algorithm technique for multiclass gene selection is introduced. The method optimizes the multiple fire flies in the multiclass type of microarray datasets to select the genes. This technique is compared with existing gene selection methods and the outcomes shows that this technique achieves high classification accuracy with less complexity than the existing methods.

4 Microarray Data

This section is devoted to the details of microarray data that are present in the literatures. Gene expression data can either be unlabeled, fully labeled or partially labeled. Data that are marked with some meaningful labels or classes is the labeled data. Unlabeled data on the other hand contain the features without the presence of labels with any explanation or information.

The selection of feature subset that make used of labeled data is called supervised feature selection. Unlabeled data are used in unsupervised feature selection and semi supervised used the semi labeled data. In this review paper, supervised feature selection is mainly involved. A labeled microarray data can be classified into two types of the dataset which are binary and multiclass. Binary data normally consists of normal and malignant tissues which have been used to separate healthy patients form cancer patients. On the other hand, multiclass data is used to distinguish different type of tumors [25] in which the classification task becomes more complicated. The works reviewed in the previous section are employed for both types of data. The type and source of data that have been used in the literatures are illustrated in Table 1. Majority of the microarray data are publicly available for the research community.

Based on Table 1, there are few works done with the same benchmark dataset. The works in [15] and [16] utilized different techniques on feature selection and classification of the same dataset. Two different objects may look-alike in an enormous feature space, therefore, [15] claimed that Euclidian distance might not be suitable for high

dimensional data. Therefore, [15] proves that implementation of the Hamming distance in feature selection managed to increase the accuracy when compared to normal Euclidean distance done in [16].

There are three works [17, 20] and [21] that used the same dataset [27] in which [20] and [21] are from the same author but with different techniques. Based on Table 1, the average accuracy for all the dataset used in [20, 21] achieved more than 90%. There is only slight difference of accuracy between the two techniques. PSO is applied in [17] with more dataset compared to [20, 21]. The work done in [17] also managed to classify the dataset to more than 90% accuracy except for Brain_Tumor1 which only managed to get 57.03% accuracy.

5 Conclusion

Classifying cancer using microarray's gene expression data is challenging because microarray has a high dimensional data and low sample dataset. Hence, many techniques have been suggested to address this problem. The ability of swarm intelligence algorithms to eliminate irrelevant genes and identify the informative genes has become a major interest among researchers. In this paper, usage of PSO, ABC, ACO and Firefly swarm intelligence algorithms in DNA microarray data analysis is reviewed. Based on the related works on feature selection, the optimization algorithms managed to improve the accuracy of classification algorithms with minimum number of selected genes for binary and multiclass type of microarray data.

Acknowledgements We would like to thank Multimedia University for their assistance and this work is supported by FRGS grant (FRGS/1/2015/TK04/MMU/03/2).

References

1. Mutingi, M., Mbohwa, C.: A Fuzzy-based particle swarm optimization algorithm for nurse scheduling. In: Proceedings of the World Congress on Engineering and Computer Science, pp. 22–24. (2014)
2. Chuang, L., Lin, Y., Yang, C., Swarm, A.P., Pso, O.: An improved particle swarm optimization for data clustering. In: Proceedings of the International MultiConference of Engineers and Computer Scientist, pp. 1–6. (2012)
3. Ali, R.S., Almousawi, A.K.: Design an optimal PID controller using artificial bee colony and genetic algorithm for autonomous mobile robot. Int. J. Comput. Appl. **100**(16), 8–16 (2014)
4. Larran, P., Saeys, Y.: Gene expression—a review of feature selection techniques in bioinformatics. Bioinformatics **23**(9), 2507–2517 (2007)
5. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science AAAS **286**(5439), 531–537 (1999)
6. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
7. Christian Blum, D.M.: Swarm Intelligence—Introduction and Applications. Springer, Berlin (2008)
8. Jérôme, L.J.D., Onwunalu, E.: Application of a particle swarm optimization algorithm for determining optimum well location and type. Comput. Geosci. **14**(1), 183–198 (2010)
9. Matekovits, L., Mussetta, M., Pirinoli, P., Selleri, S., Zich, R.E.: Improved PSO algorithms for electromagnetic optimization. In: IEEE antennas and propagation society international symposium, pp. 33–36. (2005)
10. Cristian, D., Barbulescu, C., Kilyeni, S., Popescu, V.: Particle swarm optimization techniques. Power systems applications. In: 2013 6th international conference on human system interactions (HSI), Sopot, pp. 312–319. (2013)
11. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Global Optim. **39**, 459–471 (2007)
12. Bai, Q.: Analysis of particle swarm optimization algorithm. Comput. Inf. Sci. **3**(1), 180–184 (1998)
13. Sahu, B., Mishra, D.: A novel feature selection algorithm using particle swarm optimization for cancer microarray data. Procedia Eng. **38**, 27–31 (2012)
14. Kar, S., Das Sharma, K., Maitra, M.: Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst. Appl. **42**(1), 612–627 (2015)
15. Banka, H., Dara, S.: A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation. Pattern Recogn. Lett. **52**, 94–100 (2015)
16. Sekhara, C., Annavarapu, R., Dara, S., Banka, H.: Original article: cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. EXCLI J. **15**, 460–473 (2016)
17. Chen, K., Wang, K.-J., Tsai, M.-L., Wang, K.-M., Adrian, A.M., Cheng, W.-C., Yang, T.-S., Teng, N.-C., Tan, K.-P., Chang, K.-S.: Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform. **15**, 0–9 (2014)
18. Yasodha, P., Anathanarayanan, N.R.: Analysing big data to build knowledge based system for early detection of ovarian cancer. Indian J. Sci. Technol. **8** (2015)
19. Garro, B.A., Rodríguez, K., Vázquez, R.A.: Classification of DNA microarrays using artificial neural networks and ABC algorithm. Appl. Soft Comput. **38**, 548–560 (2016)
20. Alshamlan, H.M., Badr, G.H., Alohali, Y.A.: Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Comput. Biol. Chem. **56**, 49–60 (2015)
21. Alshamlan, H., Al-Ohali, Y., Badr, G.: mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Hindawi **2015**, (2015)
22. Ganesh kumar, P., Rani, C., Devaraj, D., Victoire, T.A.A.: Hybrid ant bee algorithm for fuzzy expert system based sample classification. IEE Trans. Comput. Biol. Bioinform. **11**(2), 347–360 (2014)

23. Pal, N.S.: Robot path planning using swarm intelligence: A survey. Int. J. Comput. Appl. **83**(12), 5–12 (2013)
24. Manoharan, G.V., Shanmugalakshmi, R.: Multi-objective firefly algorithm for multi-class gene selection. Indian J. Sci. Technol. **8**, 27–34 (2015)
25. Bolon-Canedo, F.H.V., Sanchez-Marono, N., Alonso-Betanzos, A., Benitez, J.M.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
26. GEDatasets. http://sdmc.lit.org.sg/GEDatasets/. Last accessed 11 Jan 2017
27. Microarray cancer datasets. http://www.gems-system.org. Last accessed 11 Jan 2017
28. Gene Expression Datasets. http://research.nhgri.nih.gov/microarray/Supplement/. Last accessed 11 Jan 2017
29. Normalized gene expression data. http://tcga-data.nci.nih.gov/. Last accessed 11 Jan 2017
30. Leukaemia. http://www.genome.wi.mit.edu/MPR. Last accessed 11 Jan 2017
31. Lymphoma. http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt. Last accessed 11 Jan 2017
32. Colon Cancer. http://microarray.princeton.edu/oncology. Last accessed 11 Jan 2017
33. Bloomfield, C.D., Lander, E.S., Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 531–537 (1999)
34. Rosenwald, L.M.S.A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gas-coyne, R.D., Muller-Hermelink, H. K., Smeland, E.B., Giltnane, J.M., Hurt, E.M., Zhao, H., Averett, L., Yang, L., Wilson, W.H., Jaffe, E.S., Simon, R., Klausner, R.D., Powell, J., Duffey, P.L.: The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N. Engl. J. Med. **346**, 1937–1947 (2002)
35. van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A. A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Rutgers, E. T., Glash, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Bernards, R., Rutgers, E.T., Friend, S.H.: A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. **347**, 1999–2009 (2002)
36. van 't Veer, L.J., Dai, H., van de Vijver, M.J., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**, 530–536 (2002)
37. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**, 203–209 (2002)
38. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligoneu-clotide arrays. Proc. Nat'l Acad. Sci. USA **96**, 6745–6750 (1999)
39. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T. C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M.: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature **403**, 503–511 (2000)
40. National Cancer Institute Homepage. https://www.cancer.gov/. Last accessed 11 Jan 2017