Feature Selection Software Development Using Artificial Bee Colony on DNA Microarray Data

Wildan Andaru
Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
andaruwildan@gmail.com

Iwan Syarif
Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
iwanarif@pens.ac.id

Ali Ridho Barakbah
Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
ridho@pens.ac.id

DNA Microarray data is a high-dimensional data that enables the researchers to analyze the expression of many genes in a single reaction quickly and in an efficient manner. Its characteristics such as small sample size, class imbalance, and data complexity causes it difficult to classified. Feature selection is a process that automatically selects features that are most relevant to the predictive modeling in dataset. This research aims at investigating, implementing, and analyzing a feature selection method using the Artificial Bee Colony (ABC) approach. The result is compared with other evolution algorithms, which is Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). The result is that feature selection using ABC has a better result at classification using k – Nearest Neighbor (k-NN) and Decision Tree (DT), but has a slightly higher fracture of features compared to GA and PSO algorithms.

Keywords—Feature Selection, Artificial Bee Colony, Microarray Data, Data Mining

I. INTRODUCTION

Advances in bioinformatics technology such as microarray in the past few years have provided researchers with opportunities and challenges for machine learning. Microarray technology enables researchers to collects the expression levels of thousands of genes in a single experiment [1]. Data provided by microarray could be useful for diagnosis (binary approach) or distinguishing specific type of disease (multiclass approach) by input it as training data in classification. However, the number of samples of microarray data is very small (usually less than 100) compared to the number of its features (ranges from 6000 – 60,000) makes classification doesn't performs well. Having high features but small samples leads the likelihood of false prediction is high [2].

Hence, microarray data requires preprocessing techniques such as feature selection before it could be used in classification to avoid problem caused by its high features [3]. Feature selection is a preprocessing method that identifies and chooses features that are relevant and useful for analysis and predictions. Typically, there are 2 approaches in feature selection: filters and wrappers method. The main difference between these two approaches is that wrapper method using predictor algorithm as part of selection process, while a filter method doesn't, and only use the characteristic of training data to selects the features. Many swarm algorithms have been used for feature selection, such as Particle Swarm Optimization (PSO) [4]. and Ant Colony Optimization (ACO) [5]. Artificial Bee Colony (ABC) is a relatively new algorithm that proposed to solve optimization problems. ABC algorithm itself have been used in many field and proven to perform well in solving several computational problems [6, 7]. However, ABC algorithm has never been used before for feature selection on high features data and this research evaluates the performance of ABC algorithm based feature selection on high features microarray data.

II. FEATURE SELECTION

Feature selection as a preprocessing method has become indispensable to optimize machine learning performance on high-dimensional data such as DNA microarray data. Feature selection itself described as process of removing irrelevant features from the training data, so learning algorithm can focus on features that useful for analysis and prediction [8]. This process is significant, especially in high features data to improve its analysis performance. The advantages of feature selection are:

A. Avoid Overfitting

Information given by training data is too specific hence not compatible for data in general

B. Increase Accuracy

Less redundant and noise means modeling accuracy improves

C. Reduces Compotational Cost

Less features means less computational time and resources

Basically, there are 2 approaches in feature selection: filter and wrapper model [9, 10].

A. Filter

Filter approach evaluate subset of features using the data characteristics such as statistical measures, in which typically a single feature or subset of features is evaluated against the class label.

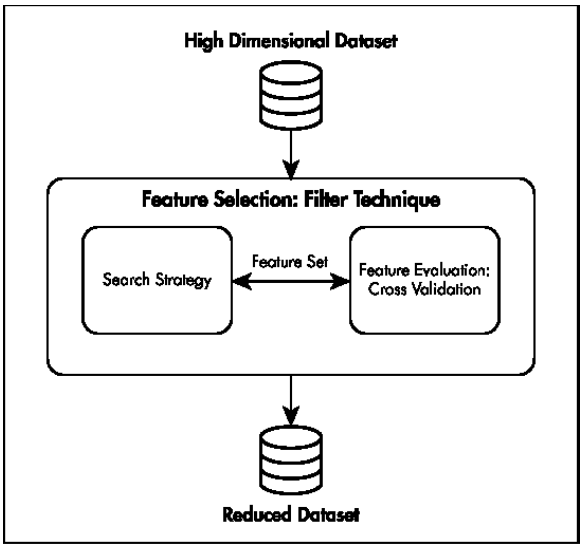

FIG. I. FILTER APPROACH

B. Wrapper

Wrapper approach generate subset by adding and removing features then evaluate it using attribute evaluator. This approach consists of 2 main components: search method and attribute evaluator. Search method searches possible subset while attribute evaluator evaluates the subset accuracy submitted by search method.

In this research, ABC algorithm used as search method in wrapper approach feature selection.

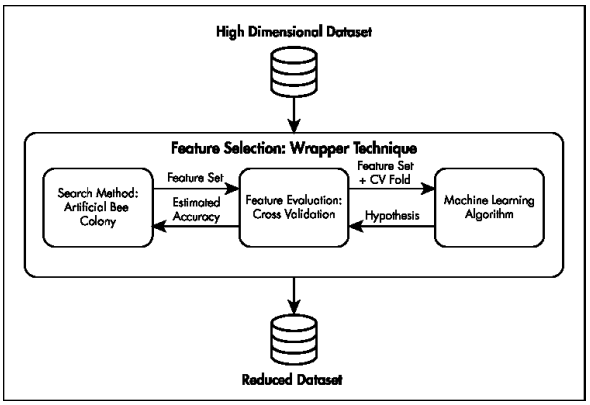

FIG. II. WRAPPER APPROACH

III. ARTIFICIAL BEE COLONY ALGORITHM

ABC is a swarm intelligent algorithm that simulating foraging activity of honey bees proposed by Karaboga in 2005 to optimize typical numerical problem [11, 12]. Basic model of ABC consists of three components: food sources, employed bees, and unemployed bees [6]:

A. Food Sources

Represents the solutions. Each food source has components such as nectar value (fitness) and its exploration limit number.

B. Employed Bees

Each employed bee exploiting particular food source. They share the food source information with unemployed bees.

C. Unemployed Bees

There are 2 types of unemployed bees: onlooker and scout. Onlooker bees receive information from employed bees about certain food sources and choose a better food sources to explore. Secondly, scout bees find a new food sources to replace abandoned food sources that has been explored so many times.

Pseudocode of basic ABC algorithm is shown below:

1. Initialize food sources
2. **Repeat**
3. **For each food sources**
4. Employed bees exploring existing food source to get new food source
5. Calculate nectar value of new food source
6. Apply greedy selection, reset trial value if new food source chosen, otherwise increment it
7. **End for each**
8. **For each food sources**
9. Onlooker bees exploring existing food source to get new food source based on nectar value of food source
10. Calculate nectar value of new food source
11. Apply greedy selection, reset trial value if new food source chosen, otherwise increment it
12. **End for each**
13. Determine abandoned food source
14. Scout bees searching new food source to replace abandoned one
15. **Until** (cycle = maximum cycle)

In basic ABC algorithm, creation of food sources based on equation 1

$$x_{ij} = x_j^{min} + rand(0,1)\left(x_j^{max} - x_j^{min}\right) \quad (1)$$

where i = 1, . . . ,N, j = 1, . . . ,F, that N is the total number of food sources and F is the total number of features.

In employed bee phase, each bee explores and find the neighbor of existing food sources associated to them. The neighbor exploration defined using equation 2

$$v_{ij} = x_{ij} + \phi_{ij}\left(x_{ij} - x_{kj}\right) \quad (2)$$

For each food source x_i, a neighbor is v_i is defined using modification. j and k is randomly generated. Value of k = 1,2, . . ., O and must be different from i. ϕ_ij is a real number between -1 and 1.

Fitness value of food source obtained through equation 3

$$fit_i = \begin{cases} \frac{1}{1 + f_i} & if\ f_i \geq 0 \\ 1 + abs(f_i) & if\ f_i < 0 \end{cases} \quad (3)$$

where f_i is an evaluation function. In some problems, the evaluation function can be directly used as a fitness value.

In onlooker bees phase, employed bees share the information about the value of food sources with onlooker bees. The probability of a food source chosen by onlooker bees to be explored defined in equation 4

$$p_i = \frac{fit_i}{\sum_{n=1}^{F} fit_i} \quad (4)$$

In scout bees phase, abandoned food that has highest trial value and higher than limit value defined, a new food source randomly generated to replace its place.

IV. DNA Microarray Data

As mentioned in introduction, typically, there are two types of microarray data, binary and multiclass. Binary datasets usually consist of healthy person samples and diseased ones while multiclass usually distinguish between different types of disease. Classifying microarray data poses several serious problems because of their characteristics.

A. Class Imbalance

This common problem happened when data dominated by a major class or classes which have significantly more samples than other class or classes. This problem causes specific rules that predict examples from minor class or classes are ignored because general rules are preferred. Therefore, minor class or cases are more often misclassified than those from major class or classes [13].

B. Small Sample

Common problem when dealing with DNA microarray data is its small sample size which is usually less than 100 instances. This is causes problem because error estimation is impacted by small samples [14].

C. Data Complexity

This characteristic describe data that difficult to classify because overlapping among classes, separability, or the linearity of the decision boundaries [15]. DNA microarray data usually has this characteristic.

V. Artificial Bee Colony Algorithm for Feature Selection

The main difference between typical optimization problems and feature selection problems is the solutions (food sources). Solutions in optimization problems are usually represented in real values while in feature selection represented in bit vector with the size of F, where F is the total number of features. Each solution has a fitness (nectar) value which is the classification accuracy of subset given by attribute evaluator.

Process of proposed method is described as follows:

- Create solutions as much as N where N is total food source defined. Initialize each solution as a bit vector with size of F, where F is the total number of features, then assign random features to each solution. Also define limit number of failed exploration, maximum cycle, and modification rate (MR).
- Submit each solution to the attribute evaluator and use the classification accuracy as fitness value of solution.
- This step is employed bees phase. Define neighbor of each food source by modify it using MR parameter. Unlike typical optimization problem, in feature selection, perturbation is performed by MR. For each feature, a random number R is generated in the range of 0 and 1. If this value is lower than MR, the feature is modified. This is expressed in equation 5:

  $$x_i = \begin{cases} modify, & R_i < MR \\ x_i, & R_i \geq MR \end{cases} \quad (5)$$

  Submit neighbor solution to the attribute evaluator and apply greedy selection between existing solution and neighbor solution. If existing solution is still better than neighbor solution, increment its trial number by 1, otherwise reset it into 0.
- Step 4 and 5 are onlooker bees phase. Calculate the probability of each solution chosen by onlooker bees using equation 6:

  $$P_i = \frac{F_i}{F_{max}} \quad (6)$$

  where F_i is the fitness value of solution and F_max is nectar value of best solution that has the highest fitness value among other solutions.
- For each solution, a random number R is generated. If probability P from step 4 is higher than R, then solution is exploited by onlooker bees. Exploitation process is the same with step 3.
- This is scout bees phase. Find abandoned solution that has the highest trial number and make sure it is higher than trial limit number that defined in initialization phase. Replace that solution with new randomly initialization solution and reset its trial number by 0.
- Repeat step 3 – 6 until requirements are met or the maximum cycle already reached.

VI. RESULT

A. Performance Mesurement

Table I shows metric used to evaluate the performance of classification.

TABLE I. PERFORMANCE MATRIX

Actual Label \ Predicted Label	*Positive*	*Negative*
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

B. Datasets

This research using 7 DNA microarray datasets which have the number of attributes from 2309 to 15155. The list of datasets is shown in table II

TABLE II. DATASETS USED

Dataset name	Sample size	Number of attributes	Source
CNS	60	7130	[16]
Leukemia	72	7130	[16]
Lung	203	12583	[16]
Lymphoma	66	4027	[16]
MLL	72	12583	[16]
Ovarian	253	15155	[16]
SRBCT	83	2309	[16]

C. Parameters

1) ABC
- Food number: 30
- Max foraging cycle: 40
- Max trial: 6
- Modification rate: 0.25
- Random seed: 1

2) GA
- Population size: 30
- Max generations: 40
- Crossover probability: 0.6
- Mutation probability: 0.033
- Random seed: 1

3) PSO
- Particle size: 30
- Max generations: 40
- C1: 1.0
- C2: 2.0
- Prune: false
- Random seed: 1

D. Performance Results

TABLE III. FEATURE SELECTION USING ABC

Dataset name	# (Before FS)	# (After FS)	Fraction of features (FF)
CNS	7130	2076	29.12%
Leukimia	7130	2542	35.65%
Lung	12583	4113	32.69%
Lymphoma	4027	797	19.79%
MLL	12583	3158	25.10%
Ovarian	15155	4690	30.95%
SRBCT	2309	621	26.89%
Average			28.60%

Table III shows that feature selection using ABC has an average fraction of features of 28.60 %. Best performance occurred in Lymphoma dataset, where ABC succeed reduced the features until reached 19.79 % from its original features. Worst performance showed in Leukemia dataset which is only succeed reduced the features until 35.65 % from its original features.

TABLE IV. FEATURE SELECTION NUMBER OF FEATURES RESULTS USING ABC, GA, AND *PSO*

Dataset name	# (Before FS)	Feature selection results (#)		
		ABC	GA	PSO
CNS	7130	2076	1755	744
Leukemia	7130	2542	2611	3173
Lung	12583	4113	5104	1530
Lymphoma	4027	797	1127	1294
MLL	12583	3158	190	2320
Ovarian	15155	4690	3225	1649
SRBCT	2309	621	170	474

Table IV shows that PSO has slightly better results in dimensionality reduction compared to ABC and GA. From 7 datasets, ABC and GA only outperform in 2 datasets, ABC in Leukemia and Lymphoma, GA in MLL and SRBCT while PSO outperform in 3 datasets, CNS, Lung, and Ovarian.

TABLE V. FEATURE SELECTION FRACTION OF FEATURES RESULTS USING ABC, GA, AND PSO

Dataset name	# (Before FS)	Feature selection results (FF)		
		ABC	GA	PSO
CNS	7130	29.12%	24.61%	10%
Leukemia	7130	35.65%	36.62%	45%
Lung	12583	32.69%	40.56%	12%
Lymphoma	4027	19.79%	27.99%	32%
MLL	12583	25.10%	1.51%	18%

Ovarian	15155	30.95%	21.28%	11%
SRBCT	2309	26.89%	7.36%	21%
Average		28.6%	22.85%	21%

Table V shows that average fraction of features in PSO also outperform ABC and GA with 21 %, while GA 22.63 % and ABC 28.60 %. The smaller the fraction of features, the better algorithm performance. However, fraction of features is not the only performance indicator of feature selection, there is still classification accuracy to evaluate feature selection performance of each algorithm.

TABLE *VI.* CLASSIFICATION *PERFORMANCE USING DECISION TREE*

Dataset name	Classification results (DT)			
	Before FS	FS ABC	FS GA	FS PSO
CNS	58.3 %	63.3 %	60 %	43.3 %
Leukemia	83.3 %	84.7 %	84.7 %	83.3 %
Lung	93.1 %	89.6 %	94.6 %	85.7 %
Lymphoma	92.4 %	87.9 %	93.9 %	87.9 %
MLL	84.7 %	76.4 %	72.2 %	81.9 %
Ovarian	95.6 %	96.4 %	96 %	95.3 %
SRBCT	84.3 %	81.9 %	81.9 %	87.9 %

Table VI shows that ABC and GA outperform PSO in classification performance using Decision Tree with each has best results in 3 datasets. ABC outperform in CNS, Leukemia (shared with GA), and Lymphoma. These results also outperform accuracy of original datasets. GA in Leukemia (shared with ABC), Lung, and Lymphoma. These results also outperform accuracy of original datasets. While PSO only in MLL and SRBCT.

TABLE VII. CLASSIFICATION PERFORMANCE USING RULE INDUCTION

Dataset name	Classification results (RI)			
	Before FS	FS ABC	FS GA	FS PSO
CNS	58.3 %	63.3 %	65 %	60 %
Leukemia	87.5 %	87.5 %	86.1 %	87.5 %
Lung	89.6 %	89.2 %	88.2 %	88.2 %
Lymphoma	93.9 %	89.4 %	84.8 %	77.3 %
MLL	91.7 %	69.4 %	83.3 %	76.4 %
Ovarian	98.4 %	96.4 %	97.6 %	96.4 %
SRBCT	86.7 %	81.9 %	87.9 %	85.5 %

Table VII shows that GA has better results in classification using Rule Induction with 4 datasets: CNS, MLL, Ovarian, and SRBCT. But doesn't outperform original datasets performance in Ovarian with difference 0.8 %. While ABC has better results in 3 datasets: Leukemia (shared with PSO), Lung, and Lymphoma. But all these results don't outperform the performance of original datasets. Lastly, PSO only outperform in Leukemia (shared with ABC) and the accuracy is same with original dataset.

TABLE VIII. CLASSIFICATION PERFORMANCE USING K-NEAREST NEIGHBORS

Dataset name	Classification results (K-NN)			
	Before FS	FS ABC	FS GA	FS PSO
CNS	51.7 %	63.3 %	56.7 %	58.3 %
Leukemia	87.5 %	91.7 %	93.1 %	91.7 %
Lung	91.1 %	90.1 %	91.1 %	91.6 %
Lymphoma	97 %	98.5 %	97 %	97 %
MLL	80.6 %	81.9 %	75 %	87.5 %
Ovarian	95.6 %	96.4 %	95.6 %	95.6 %
SRBCT	83.1 %	96.4 %	84.3 %	86.7 %

Table VIII shows that accuracy of ABC-reduced datasets outperforms other algorithms with 4 datasets: CNS, Lymphoma, Ovarian, and SRBCT. These results also outperform performance of original datasets. Secondly, PSO has best results in 2 datasets: Lung and MLL. These results also outperform performance of original datasets. Lastly GA only has best results in Leukemia which is also outperform original dataset.

VII. CONCLUSION

In terms of dimensionality reduction, GA and PSO shows a slightly better result than ABC. PSO has best average fraction of features with 21 %, while GA average is 22.85 % ABC average is 28.60 %. However, fraction of features is not the only performance indicator of feature selection. Classification accuracy also plays a big role in this case. Classification accuracy of ABC-reduced datasets shows a tie results with GA-reduced datasets, and better than PSO-reduced datasets. GA-reduced datasets have best results in classification using k-Nearest Neighbor, which is better than GA and PSO. And in Decision Tree classification shows a same result with GA, and better than PSO. These results also almost outperform the classification performance of original datasets. Overall, ABC has same performance in feature selection with other evolutionary algorithms.

REFERENCES

[1] D. Shalon, S. J. Smith and P. O. Brown, "A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization," *Genome Research,* vol. 6(7), pp. 639-645, 1996.

[2] G. Piatetsky-Shapiro and P. Tamayo, "Microarray Data Mining: Facing the Challenges," *ACM SIGKDD Explorations Newsletter,* vol. 5, no. 2, pp. 1-5 , 2003.

[3] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Transactions On Pattern Analysis And Machine Intelligence,* vol. 19, no. 2, pp. 153 - 158, 1997.

[4] X. Wang , J. Yang, X. Teng, W. Xia and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters ,* vol. 28, p. 459 – 471, 2007.

[5] Z. Yan and C. Yuan, "Ant colony optimization for feature selection in face recognition," in *In: Zhang D., Jain A.K. (eds) Biometric Authentication. Lecture Notes in Computer Science, vol 3072*, Springer, Berlin, July 15–17 2004.

[6] D. Karaboga, B. Gorkemli, C. Ozturk and N. Karaboga, "A Comprehensive Survey: Artificial Bee Colony (ABC) Algorithm and Applications," *Artificial Intelligence Review,* vol. 42, no. 1, pp. 21-57, June 2014.

[7] B. Basturk and D. Karaboga, "On The Performance of Artificial Bee Colony (ABC) Algorithm," *Applied Soft Computing,* vol. 8, no. 1, pp. 687 - 697, January 2008.

[8] I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh, "Feature Extraction: Foundations and Applications," *Springer,* vol. 207, 2006.

[9] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis,* vol. 1, pp. 131-156, 3 1997.

[10] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157 - 1182, 3 2003.

[11] D. Karaboga, "An Idea Based on Honey Bee Swarm For Numerical Optmization," Kayseri, 2005.

[12] D. Karaboga and B. Basturk, "A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony Algorithm," *Journal of Global Optimization,* vol. 39, pp. 459-471, April 2007.

[13] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews,* vol. 42, no. 4, pp. 463 - 484, 2012.

[14] E. R. Dougherty, "Small Sample Issues for Microarray-based Classification," *Comp Funct Genom ,* no. 2, p. 28–34, 2001.

[15] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions On Pattern Analysis And Machine Intelligence,* vol. 24, no. 3, pp. 289 - 300, 2002.

[16] Z. Zhu, Y. S. Ong and M. Dash, "Markov Blanket-Embedded Genetic Algorithm for Gene Selection," *Pattern Recognition,* vol. 49, no. 11, pp. 3236-3248, 2007.