A comparison between the wrapper and hybrid methods for feature selection on biology Omics datasets

Chirui Guo\textsuperscript{1,a} and Qiuhan Li\textsuperscript{2,b}

\textsuperscript{1}School of Science, The Hong Kong University of Science and Technology, Hong Kong, China
\textsuperscript{2}College of Urban and Environmental Science, Northwest University, Xi’an, China
\textsuperscript{a}Email: cguoaj@connect.ust.hk, \textsuperscript{b}Email: lilirenabc@gmail.com

Abstract. Classification problem in biological Omics data has gained in popularity in recent years. In consideration of the high dimension of the Omics dataset, the importance of feature selection technology becomes apparent. Feature selection can not only improve the classification accuracy but also avoid overfitting which is a common issue in machine learning. There are three main categories of feature selection methods: filter methods, wrapper methods, and hybrid methods. Yet it is difficult for researchers to choose among them. In this study, we conducted a comprehensive comparison for filter methods, wrapper methods, and hybrid methods. Specifically, we selected information gain as the filter method, genetic algorithm, and binary particle swarms optimization as the wrapper methods, IG-GA and IG-BPSO as the hybrid methods for comparison. The experimental results show that the IG-BPSO, a hybrid method, has the highest classification accuracy among these feature selection methods on three Omics datasets.

1. Introduction

Biological Omics study is a set of various disciplines including genomics, transcriptomics, proteomics, metabolomics, and glycomics. Omics studies focus on the characterization and quantification of a collection of biomolecules from one layer of the genetic central dogma.

Omics produces high-dimensional data prone to be noisy. Machine learning technology deals with high dimensional data very efficiently. Machine learning (ML) technology is always adopted in the field of cancer research for classification tasks. A big challenge in the area is that Omics data usually have a larger number of features compared to the sample size. We know there is a positive relationship between the accuracy of the classification model and the feature size at some level [1]. Moreover, large feature size may cause overfitting problems [2]. To this end, feature selection is always applied to reduce the number of features removing irrelevant features and help to improve classification performance.

There are mainly three types of feature selection methods [3]: (1) Filter Methods (2) Wrapper Methods (3) Hybrid Methods. The filter methods rank features based on specific criteria and select features through setting a threshold for these criteria [4]. The wrapper methods follow a greedy search approach by integrating a predefined classifier to evaluate all the possible combinations of features. The better performance of the predefined classifier is, the better the feature selection scheme is. The feature selection schemes become better and better after several iterations. Hybrid methods combine filter methods and wrapper methods. On the first stage, features are selected using filter methods. On the second stage, features are further selected by wrapper methods from the features of the first stage.
Various feature selection methods have been proposed previously. Battiti et al. introduced a filter method for feature selection in which the mutual information was chosen as the criterion for evaluating features [5]. Although filter methods are easy to implement, they cannot always achieve satisfactory results. Therefore, a new wrapper approach which was called genetic algorithm was proposed for feature selection by Raymer et al. [6], and they found that the genetic algorithm has high accuracy and needed fewer features to achieve this accuracy. Aličković et al. used the wrapper method: the genetic algorithm for feature selection on breast cancer datasets which has shown new opportunities in diagnosing breast cancer [7] and found that the genetic algorithm can rank significant features correctly since the genetic algorithm performs well in terms of classification accuracy. Kennedy et al. proposed a discrete vision of particle swarms optimization called binary particle swarm optimization, which makes PSO suitable for feature selection [8]. Vieira et al. proposed a new fitness function for evolutionary computation algorithms which minimizes two objectives: the number of features and the classification error [9-10].

Because of the high time complexity of wrapper methods, Chuang et al. introduced two hybrid methods of feature selection, IG-GA, and IG-BPSO, for DNA microarray data, and obtaining a higher classification accuracy on microarray datasets compared to other feature selection methods [11-12]. It’s difficult to make a decision to select a suitable feature selection method for Omics data. In this paper, we compared one filter method, that is information gain method, two wrapper methods, genetic algorithm (GA) and binary particle swarms optimization (BPSO) on Omics datasets. Our results show that hybrid methods highly improved the classification accuracy comparing to filter methods and wrapper methods, and IG-BPSO in hybrid methods performs best on Omics datasets.

This paper is organized as follows. In Section 2, we introduce all three categories of methods for comparison. For filter methods, we selected Information Gain (IG) as our criterion. For two wrapper methods, genetic algorithm (GA) and binary particle swarms optimization (BPSO) were selected. For two classic hybrid methods, we selected IG-GA and IG-PSO for feature selection. We show the result of the comparison in Section 3 and give our conclusion in Section 4.

2. METHODS

2.1. Datasets

We used three Omics datasets in this study. Leukaemia dataset [13], Mixed-lineage leukaemias dataset (MLL) [14], and diffuse large B-cell lymphoma dataset (DLBCL) [15]. They can be downloaded respectively from:

- http://portals.broadinstitute.org/cgi-bin/cancer/publications/view/43
- http://portals.broadinstitute.org/cgi-bin/cancer/publications/view/63
- http://leo.ugr.es/elvira/DBCRepository/DLBCL/DLBCL-Stanford.html

Table 1 gives the sample size, the number of features and classes of the three datasets.

| Datasets  | Samples | Features | Classes |
|-----------|---------|----------|---------|
| Leukaemia | 72      | 7129     | 2       |
| MLL       | 72      | 12582    | 3       |
| DLBCL     | 47      | 4026     | 2       |

- The Leukaemia dataset consists of 72 samples and 7129 gene expression levels. There are two categories in this dataset which are Acute Myeloid Leukaemia (AML) and Acute Lymphoblastic Leukaemia (ALL). This dataset contains 46 ALL and 26 AML cases.
- The MLL dataset consists of 72 samples and 12582 gene expression levels. There are three classes in the dataset: Acute Myeloid Leukemia (AML), Acute Lymphoblastic Leukemia (ALL), and Mixed-lineage leukaemias (MLL). 28 AML, 24 ALL, and 20 MLL cases are contained in the dataset. It
is a multiclass problem for classification. We decided to use the One-versus-Rest method to solve this problem. One-versus-Rest method trains Binary Classifier models for any class as positive samples and all other samples as negatives. In this study, three binary classifiers need to be trained, and we can obtain three scores for each prediction, then we choose the prediction with the highest score [16].

- Diffuse large B-cell lymphoma dataset (DLBCL) consists of 47 samples and 4026 gene expression levels. There are two classes in this dataset which are 'germinal centre B-like' group and 'activated B-like' group. This dataset contains 24 'germinal centre B-like’ samples and 23 ‘activated B-like’ samples.

2.2. Information Gain
IG is also called Kullback–Leibler divergence or relative entropy. In this study, IG is a synonym for mutual information [17]. It is a filtering criterion for feature ranking technique which is widely used in filter method for feature selection. Features are ranked using IG, and then a threshold needs to be set to choose the features above this threshold, if not set a threshold, we can select the top K features.

To describe IG, we must start with the definition of entropy. In probability theory, the entropy of a random variable measures the uncertainty about the value that might be assumed by the variable. For Omics datasets, entropy represents the uncertainty of the whole dataset itself. It can be written as equation (1) [5]:

\[
H(Y) = -\sum_{i=1}^{n} P(y_i) \log P(y_i)
\]  

where \(Y\) denotes a set of training examples with possible values \(\{y_1, y_2, \ldots, y_n\}\), \(P(Y)\) denotes the probability mass function.

Then Shannon defined the conditional entropy of two events, \(X\) and \(Y\) taking values \(x_i\) and \(y_j\) respectively, as equation (2):

\[
H(Y|X) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)}
\]

On Omics datasets, if \(Y\) denotes all of the variables and \(X\) denotes one specific gene, conditional entropy measures the uncertainty of the dataset \(Y\) on the condition of a specific variable \(X\) taking a certain value \(x\). Furthermore, the IG is the decrease in uncertainty, and it also measures the importance of a specific feature \(X\) for an Omics dataset \(Y\). It can be written as equation (3):

\[
I(Y, X) = H(Y) - H(Y|X)
\]

IG with a larger value indicates more information is contained in the corresponding feature.
2.3. Genetic Algorithm

Genetic algorithm is one of the evolutionary algorithms, and it is also a kind of stochastic search algorithm. This algorithm is inspired by the Darwinian principle. It is widely used to generate high-quality solutions to optimization, search problems, and machine learning problems. After several iterative, the global maximum for the objective function can be found eventually [18]. The GA consists of six important steps: Initialization, Crossover, Mutation, Fitness evaluation, Selection, Termination. Figure 1 shows the workflow of the genetic algorithm.

Initialization/Updating: The algorithm starts with a population of candidate solutions that mimic the structure of the chromosome. A chromosome represents a selecting scheme and it is represented as a binary vector where 1 indicates the feature will be selected and 0 indicated the feature will be eliminated. Moreover, the population in each iteration is called a generation. and at the beginning of the algorithm we usually randomly generate each chromosome. We update these chromosomes until termination conditions are met.

Crossover: The crossover imitates the behaviour of Chromosome in the natural world. It combines two parents to form children for the next generation. At first, we should randomly choose some crossover points for each pair of parents. and swap the genes of this pair of parents between these crossover points. Then we get a range of new offspring.

Mutation: The mutation randomly impacts the properties of the offspring, which provides diversity for the new population. For binary encoded GAs, we randomly choose one or more genes in a chromosome and flip them. If the selected gene is 1, we change it to 0, and if the selected gene is 0, we change it to 1. Both mutation and crossover bring a considerable complication to successive generations.

Figure 1. Genetic Algorithm
Fitness evaluation: After mutation, we can evaluate each new chromosome through a specific function that can map a binary vector into a real number. The higher the value is, the better the chromosome is. There are many fitness functions, and we chose the fitness function as below [9-10]:

\[ f(u_i) = \alpha(1 - P) + (1 - \alpha) \left( 1 - \frac{N_f}{N_t} \right) \]  

(4)

where \( P \) is the classifier performance measure, \( N_f \) is the size of the tested feature subset, and \( N_t \) denotes the size of the total feature set.

Selection: Inspired by the Darwinian principle: 'It is not the strongest of the species that survive, but the one most responsive to change', this algorithm compares the fitness values which generates from the fitness evaluation step, and select the top \( N \) chromosomes or schemes of the corresponding fitness value as the new population, from the pool of parents and offspring. Chromosomes with high fitness value have more chance to be selected as the next candidate solutions.

Termination: The generational process is stopped when a termination condition has been reached. Empirically we set a certain number of generations as algorithms’ termination, or Genetic Algorithm finds a solution that satisfies minimum criteria. In this paper, we choose 100 generations as termination.

2.4. Binary Particle Swarms Optimization

Particle Swarm Optimization (PSO) inspired by the behaviours of birds’ flock is a common stochastic approach for continuous optimization tasks. It is one of the swarm intelligence techniques. In PSO, we call each potential solution a particle, and each particle has two properties: position and velocity. In the \( t \)th iteration, the position of particle \( i \) is represented by a vector \( x^i_t = (x^i_1,x^i_2,...,x^i_d) \). Particles move in the search space for a better solution by changing their velocity, which can also be represented by a vector \( v^i_t = (v^i_1,v^i_2,...,v^i_d) \). The updating rules of velocity can be written as the following equations:

\[
\begin{align*}
    v^+_{ij} &= \omega v^i_{ij} + c_1 \times \text{rand}_1 \times (pbest_{ij} - x^i_{ij}) \\
    &\quad + c_2 \times \text{rand}_2 \times (gbest_{ij} - x^i_{ij}) \\
\end{align*}
\]

(5)

where \( \omega \) is inertia weight, \( c_1 \) and \( c_2 \) are acceleration constants, \( \text{rand}_1 \) and \( \text{rand}_2 \) are random values generated from a uniform distribution. \( pbest_i = (pbest_{i1},pbest_{i2},...,pbest_{ij}) \) represents the best previously encountered position of the \( i \)th particle. The global best position of all the population is called \( gbest = (gbest_1,gbest_2,...,gbest_m) \), so \( pbest_{ij} \) and \( gbest_{ij} \) are elements in \( j \) dimension of \( gbest \) and \( pbest_i \). Then the updating rule of a position is given according to the following equation:

\[
x^+_{ij} = x^i_{ij} + v^i_{ij}
\]

(6)

Binary particle swarms optimization (BPSO) is a discrete version of PSO to solve discrete optimization tasks. For Omics datasets, we can encode each scheme of feature selection as a pool of binary series, which allows us to select features using BPSO [19]. However, the updating rule of BPSO must be modified based on PSO. The position in BPSO consists of \([0,1]\), which means ‘selected’ and ‘unselected’. A sigmoid function is applied in BPSO to transform \( v^i_{ij} \) into a probability within the range of \((0,1)\), and it can be written as the following equation:

\[
s(v^i_{ij}) = \frac{1}{1 + e^{v^i_{ij}}}
\]

(7)

the updating rule is modified to the following equation:

\[
x^+_{ij} = \begin{cases} 
1, & \text{if } \text{rand()} < s(v^i_{ij}), \\
0, & \text{otherwise} 
\end{cases}
\]

(8)
where the \( \text{rand()} \) is also a random value generating from a uniform distribution within the range of \((0,1)\). Equation (4) was chosen as the fitness function when BPSO applied to feature selection.

2.5. **IG-GA and IG-BPSO**

![Workflow of hybrid methods](image)

IG-GA is a hybrid method for feature selection that combines the filter method (IG) with a wrapper method (GA). In the first stage, we calculate IG for each gene, and then set a threshold to choose genes whose information gain is higher than the threshold. We can also select a definite number of genes whose IG is top \( k \). In the second stage, we apply GA on the data which were chosen by the first stage, and further reduce the number of selected features [12]. IG-BPSO is also a hybrid method for feature selection which combines IG with BPSO and it has the same workflow as the IG-GA method. The workflow is presented in figure 2 [10, 19].

2.6. **Evaluation**

In order to evaluate the performance of each feature selection method, we compared each method. To conduct a comprehensive comparison, three classical classifiers were fitted using features selected by different feature selection methods. They are random forest (RF), K-nearest neighbour (KNN), and support vector machine (SVM).

The performance of each method was evaluated by K-Fold cross-validation, and we set \( K = 10 \) in this study. In 10-fold cross-validation, the dataset was randomly split into ten parts. and we chose any part as the test set and let the nine parts be training set. So we can obtain ten classification accuracies. The average of 10 classification accuracies was regarded as the final accuracy of the machine learning algorithm which can be used in the fitness function.

The parameters of every algorithm are as follows.

2.6.1. **Parameters for Wrapper and Hybrid methods**

- **GA**: A Python package called DEAP was used in implementing GA. We set crossover probability equal to 1.0, mutation probability equal to 0.1, and population size was set to 30 then we chose 100 as the number of generations.
• BPSO: A Python package called Pyswarms was used in implementing BPSO. The number of particles used was 30, the two acceleration factors are both equal to 2, and the inertia weight is set to 1.0, then we chose 100 as the number of generations as GA.
• Fitness Function: We choose 0.1 and 0.9 as alpha’s value for the fitness function in both BPSO and GA, which decides the tradeoff between the classifier performance and the number of selected genes. The equation was introduced in equation (4).

2.6.2. Parameters for Classifiers
• All the Classifiers are implemented by scikit-learn, a Python package.
  – RF: We built ten trees in the forests and chose Gini Impurity to measure the quality of the split.
  – KNN: Euclidean distance was used in this study. We set the parameter \( K = 1 \), and all points were weighted equally.
  – SVM: We selected Radial Basis Function as kernel function and set the Regularization parameter \( C \) to 1.0.

3. Results
We compared the performance of each feature selection method, including a filter method: IG, two wrapper methods: GA and BPSO, two hybrid methods: IG-GA and IG-BPSO. Moreover, we added a classification performance without feature selection. In our experiment, we focus on the results of two aspects of feature selection: 1. the number of selected genes 2. classification performance of feature selection methods.

We also choose two different values of \( \alpha \) which are 0.1 and 0.9. In equation (4) we can easily know that the first part of the fitness equation is the residual error rate of the model and the second part of the fitness equation is the unselected rate of the genes. \( \alpha \) indicates the weight of the residual error rate part. It means that the larger the \( \alpha \), the more the model focuses on accuracy; conversely, the smaller the \( \alpha \), the more the model focuses on the number of features.

3.1. Number of selected genes

| Datasets | Classifier | No FS | Filter | Wrapper | Hybrid |
|----------|------------|-------|--------|---------|--------|
|          |            |       | IG     | GA      | BPSO   | IG-GA  | IG-BPSO |
| Leukaemia| RF         | 7129  | 500    | 3158    | 3558   | 147    | 250     |
|          | KNN        | 7129  | 500    | 3185    | 3584   | 164    | 262     |
|          | SVM        | 7129  | 500    | 3279    | 3578   | 153    | 238     |
| MLL      | RF         | 12582 | 500    | 5764    | 6318   | 150    | 265     |
|          | KNN        | 12582 | 500    | 5722    | 6282   | 152    | 249     |
|          | SVM        | 12582 | 500    | 6103    | 6201   | 204    | 266     |
| DLBCL    | RF         | 4026  | 500    | 1737    | 2011   | 137    | 251     |
|          | KNN        | 4026  | 500    | 1679    | 2008   | 154    | 260     |
|          | SVM        | 4026  | 500    | 1752    | 2045   | 159    | 253     |

| Datasets | Classifier | No FS | Filter | Wrapper | Hybrid |
|----------|------------|-------|--------|---------|--------|
|          |            |       | IG     | GA      | BPSO   | IG-GA  | IG-BPSO |
| Leukaemia| RF         | 7129  | 500    | 3158    | 3642   | 147    | 283     |
|          | KNN        | 7129  | 500    | 3227    | 3688   | 145    | 274     |
|          | SVM        | 7129  | 500    | 3279    | 3604   | 153    | 283     |
|          | RF         | 12582 | 500    | 5829    | 6450   | 125    | 272     |
Due to the difference of feature selection strategies for feature selection methods, a different number of features will be selected. The selected gene number was manually set to 500 in IG when using the filter method, while in other methods it is determined by the algorithm itself which are all stochastic approaches, and thus different gene numbers are selected every time we run these algorithms. However, as shown in table 2 and table 3 we found that the higher value of $\alpha$ is, the fewer genes were selected, which can be explained well by the definition of $\alpha$ in equation (4): the trade-off between the classifier performance and the number of selected genes.

By comparing table 2 and table 3, we can get the following findings:

In wrapper methods, both GA and BPSO approximately cut the gene numbers in half compared to the raw Omics data, but the feature size is still too large. Overall, GA performs better than BPSO in terms of the number of remaining features (number of genes).

In hybrid methods, it is apparent that hybrid methods can remarkably reduce the data dimension, which highly improved the wrapper methods’ ability to reduce features. It will also reduce the cost of computing hours. The result is similar to the wrapper methods, IG-GA also has a more significant gene numbers reduction than IG-BPSO, even though IG-GA has a lower classification than IG-BPSO.

Table 4. Classification accuracy (%) $\alpha = 0.9$

| Datasets | Classifier | No FS | Filter | Wrapper | Hybrid |
|----------|------------|-------|--------|---------|--------|
| Leukaemia | RF         | 78.57 | 91.61  | 85.71   | 94.46  | 94.29  | 98.57 |
|          | KNN        | 91.43 | 88.57  | 91.43   | 97.14  | 86.07  | 97.14 |
|          | SVM        | 78.57 | 92.86  | 78.57   | 88.57  | 92.86  | 97.14 |
|          | RF         | 77.68 | 91.61  | 83.04   | 92.86  | 87.32  | 97.14 |
| MLL      | KNN        | 91.61 | 94.46  | 91.61   | 97.32  | 91.79  | 98.57 |
|          | SVM        | 75.89 | 93.04  | 75.89   | 84.29  | 85.89  | 98.57 |
|          | RF         | 69.50 | 80.50  | 71.50   | 89.50  | 89.00  | 98.00 |
| DLBCL    | KNN        | 70.00 | 87.50  | 70.00   | 76.50  | 85.50  | 96.00 |
|          | SVM        | 74.00 | 94.00  | 74.00   | 81.00  | 91.50  | 100.00 |

Table 5. Classification accuracy (%) $\alpha = 0.1$

| Datasets | Classifier | No FS | Filter | Wrapper | Hybrid |
|----------|------------|-------|--------|---------|--------|
| Leukaemia | RF         | 78.57 | 90.18  | 85.71   | 91.79  | 93.04  | 98.57 |
|          | KNN        | 91.43 | 88.57  | 91.43   | 92.85  | 90.18  | 90.18 |
|          | SVM        | 78.57 | 92.86  | 78.57   | 88.57  | 92.86  | 92.86 |
|          | RF         | 77.68 | 92.86  | 83.04   | 87.14  | 93.04  | 94.29 |
| MLL      | KNN        | 91.61 | 94.46  | 91.61   | 94.46  | 94.46  | 94.46 |
|          | SVM        | 75.89 | 93.04  | 75.89   | 78.75  | 93.04  | 91.61 |
|          | RF         | 69.50 | 75.50  | 69.50   | 89.00  | 89.50  | 93.50 |
| DLBCL    | KNN        | 70.00 | 87.50  | 70.00   | 70.00  | 85.00  | 81.50 |
|          | SVM        | 74.00 | 96.00  | 74.00   | 80.50  | 93.50  | 96.00 |
3.2. Classification performance of feature selection methods
The classification accuracies of IG-GA, IG-BPSO, and several competing algorithms are shown in the table 4, and table 5 which correspond to $\alpha = 0.9$ and $\alpha = 0.1$. A higher $\alpha$ value means more weight on the accuracy of the model, as explained at the beginning of 3. Overall, compared with no feature selection, feature selection can significantly improve the accuracies of classification algorithms.

As shown in table 4 and table 5, the filter methods have greatly improved the accuracy of the model already compared to no feature selection. In DLBCL for SVM when $\alpha = 0.1$, the accuracy even increased by 24%. Wrapper methods sometimes achieve better performance than filter methods, but sometimes not. It is one of the reasons why considering about the hybrid method combining the filter method and the wrapper method. In wrapper methods, BPSO obtains higher classification accuracy than GA generally. As for hybrid methods, they obtain better performance than other competing algorithms in both three Omics datasets most of the time. That means hybrid methods that combine the filter methods and wrapper methods can not only simplify the classification problem but also improve not only the classification accuracy. In Hybrid methods, better performance was obtained using IG-BPSO compared to IG-GA.

4. Conclusion
In this work, we compared five methods: one filter method: IG, two wrapper methods: GA and BPSO, two hybrid methods: IG-GA and IG-BPSO for feature selection on three Omics datasets. Classification accuracy was chosen as our judgment standard in this study. The results show that the classification accuracy was highly improved by applying hybrid methods comparing to filter methods and wrapper methods. What is more, though GA and IG-GA can reduce the dimension more than BPSO and IG-BPSO, BPSO and IG-BPSO still have an advantage for boosting classification accuracy. To sum up, we confirm that IG-BPSO in hybrid methods performs better than other methods on Omics datasets.

Acknowledgments
I would like to express my thanks to Zilin Fan for his guidance, and also for his helpful comments on the draft of this paper.

References
[1] Janecek, A., Gansterer, W. N. W., Demel, M., Ecker, G. (2008) On the Relationship Between Feature Selection and Classification Accuracy. PMLR, 4: 90–105.
[2] Schneider, M. V., Orchard, S. (2011) Omics technologies, data and bioinformatics principles. Bioinformatics for omics Data, 3-30.
[3] Chandrashekar, G., Sahin, F. (2014) A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28.
[4] Ammu, P. K., Preeja, V. (2013) Review on feature selection techniques of DNA microarray data. International Journal of Computer Applications, 61(12).
[5] Battiti, R. (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 5(4): 537-550.
[6] Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., Jain, A. K. (2000) Dimensionality reduction using genetic algorithms. IEEE transactions on evolutionary computation, 4(2): 164-171.
[7] Aličković, E., Subasi, A. (2017) Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4): 753-763.
[8] Kennedy, J., Eberhart, R. C. (1997) A discrete binary version of the particle swarm algorithm. In: 1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation. Orlando. pp. 4104-4108.
[9] Vieira, S. M., Sousa, J. M., Runkler, T. A. (2010) Two cooperative ant colonies for feature selection using fuzzy models. Expert Systems with Applications, 37(4): 2714-2723.
[10] Vieira, S. M., Mendoça, L. F., Farinha, G. J., Sousa, J. M. (2013) Modified binary PSO for
feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing, 13(8): 3494-3504.

[11] Chuang, L. Y., Ke, C. H., Yang, C. H. (2016) A hybrid both filter and wrapper feature selection method for microarray classification. arXiv preprint arXiv:1612.08669.

[12] Yang, C. H., Chuang, L. Y., Yang, C. H. (2010) IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. Journal of Medical and Biological Engineering, 30(1): 23-28.

[13] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439): 531-537.

[14] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Korsmeyer, S. J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1): 41-47.

[15] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769): 503-511.

[16] Aly, M. (2005) Survey on multiclass classification methods. Neural Netw, 19: 1-9.

[17] Vergara, J. R., Estévez, P. A. (2014) A review of feature selection methods based on mutual information. Neural computing and applications, 24(1): 175-186.

[18] Vafaie, H., De Jong, K. A. (1992) Genetic Algorithms as a Tool for Feature Selection in Machine Learning. In: ICTAI. Arlington. pp. 200-203.

[19] Chuang, L. Y., Chang, H. W., Tu, C. J., Yang, C. H. (2008) Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 32(1): 29-38.