A Hybrid Feature Selection Algorithm Applied to High-dimensional Imbalanced Small-sample Data Classification

Fang Feng*†, Qingquan Lv‡*, Mingsong Wang‡, Xuhui Yang*, Qingguo Zhou*, Rui Zhou *
Qingguo Zhou is the corresponding author
* School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China
E-mail: zhouqg@lzu.edu.cn, fengf15@lzu.edu.cn, yangxh16@lzu.edu.cn, zr@lzu.edu.cn
† School of Electronic and Information Engineering, Lanzhou Institute of Technology, Lanzhou, Gansu, China
‡ State Grid Gansu Electric Power Research Institute, Lanzhou, Gansu, China
E-mail: lvqingquan_lzu@126.com, wmsfine@163.com,

***Abstract*—With the rapid development of microarray technology and interdisciplinary science, it is possible for microarray technology to be used to predict diseases. Microarray technology has the advantages of high speed, high efficiency and reliability in disease prediction. However, microarray data are usually high-dimensional with small samples, additionally, the samples are often imbalanced, which brings a lot of difficulties to researchers. In view of the above problems, it is proposed in this paper a Filter-Wrapper hybrid feature selection algorithm Union Information Gini Cost-sensitive Feature Selection General Vector Machine (UIG-CFGVM) to tackle the high-dimensional imbalanced small-sample problem. The improved hybrid algorithm is as follows: Firstly, the most common features are removed by the proposed hybrid filter algorithm UIG, which is obtained by Information Gain (Info)and Gini Index (Gini). Secondly, Cost-sensitive Feature selection General Vector Machine (CFGVM) is used as Wrapper method to further improve the performance of the algorithm. The experimental results show that the proposed algorithm UIG-CFGVM has better classification performance in seven biomedical high-dimensional imbalanced small-sample datasets compared with other similar algorithms.**

***Index Terms*—Filter algorithm, Wrapper algorithm , Feature selection, High-dimensional Imbalanced Small-sample data**

I. Introduction

In the past decades, cancer has become one of the serious diseases to the health of our people with the increase of its incidence and mortality. In the medical diagnosis of cancer, the classification of several tumor types is very important. Accurate prediction of tumor type can provide better treatment for patients, and if cancer can be accurately predicted in the early stage, the survival rate can be greatly improved. Traditional treatments are based primarily on the morphological features of tumor tissue. According to relevant reports, these conventional methods have some limitations. Therefore, effective methods for distinguishing cancer subtypes are essential [1].

At present, the study of gene sequence has become an important method for early cancer prediction and prevention. Usually, the research of gene sequence is mainly based on DNA microarray data. DNA microarrays are a group of tiny spots attached to solid surfaces to measure gene expression levels. The technology allows researchers to study a large number of genes, so the use of gene expression profiles for cancer diagnosis has progressed very rapidly [2]. However, gene expression profiles are usually high-dimensional with small samples, additionally, the samples are often imbalanced. As a preprocessing step of microarray data processing, feature selection has rapidly become an indispensable part of researchers. Feature selection not only removes redundant and irrelevant features, but also helps biologists to connect basic expressions with diseases [3].

In this paper, a novel algorithm UIG-CFGVM is proposed to tackle the high-dimensional small sample and imbalanced classification problems. To our best knowledge, there is no recent research that applied the GVM and BALO in the the high-dimensional small sample and imbalanced classification problems. In the high-dimensional small sample and imbalanced classification, there are not only the ”dimensional curse”, the classifier bias problem, but also the over-fitting problem, and the feature selection can not only reduce the dimension of the data, but also improve the generalization ability of the algorithm. Therefore, it has been widely studied by scholars. The general idea of the proposed algorithm is as follows: The first step is to propose an improved Filter method UIG in which the existing Filter algorithm Info and the Gini are mixed, the first n features obtained by the two algorithms are combined. In this way, the advantages of the two Filter methods can be merged without missing important features, and the second step uses the Wrapper algorithm, which uses the proposed CFGVM algorithm. The feature can be further screened and the imbalance and over-fitting problems solved to improve classification performance.

II. The related work of feature selection in high-dimensional small samples

Feature selection is an important technique for data preprocessing [4]. According to different search strategies, feature selection methods can be divided into Filter feature selection algorithm, Wrapper feature selection algorithm and Embedded feature selection algorithm.

Filter feature selection algorithm mainly depend on the general statistical features of training data without using

any learning algorithm. Hoque et al. [5] compute feature class fuzzy mutual information based on fuzzy mutual information, realize feature selection of FMIFS-ND, and select the feature with the highest mutual information. Raza et al. [6] proposed a new concept of "enhanced dependency class". IDC can replace dependency measurement and improve classification performance by reducing execution time and runtime memory. Guo et al. [7] proposed a regularized logistic regression (RLR) with support vector machine as the selection mechanism. This algorithm provides a global optimal solution with linear complexity and is superior to other feature selection algorithms.

Wrapper feature selection algorithm uses evolutionary strategy to guide search. At present, the related Wrapper algorithms include particle swarm optimization (PSO) [8], artificial bee colony algorithm [9], ADSRPCL-SVM, Ant Colony Optimization Algorithm (ACO) [10], genetic algorithm based SVM algorithm [11] and gene programming algorithm [12]. Sharma et al. [13] implements the continuous feature selection (SFS) method, which allows the size of features less than 10 to be processed at a time, and the level assigned by the features to be increased. At each stage, a feature is deleted until the stop condition is reached. Kang et al. [14] achieves global optimization by combining random forward search to select relevant features. It also uses other sequential selection techniques, such as sequential forward selection (SFS) and sequential backward elimination (SBE), until the final model is built. Separate use of SFS and SBS is vulnerable to nesting benefits [15]. In SBS, once a feature is ignored, it can not be re-selected; in SFS, once a feature is selected, it can not be deleted later. In order to mitigate these effects, some literatures use Sequential Floating Forward Selection (SFFS) and its improved extended algorithm [16, 17] to improve the quality of selected features.

Embedded feature selection algorithm, the process of learning classifier and feature selection are carried out simultaneously. For example, random forest [18] based on genome data analysis, convergent random forest [19] for drug response, and artificial neural network algorithm [20] for improving preconditional microRNA classification. Zhu et al. proposed a feature selection criterion based on the detailed analysis of multi-criteria linear programming (MCLP) classification algorithm, and designed an embedded candidate feature selection program for MCLP [21]. Mishra et al. proposed a strategic gene selection algorithm, SVM-BT-RFE, which is a variant of SVM-RFE and SVM-T-test. The algorithm takes into account the results of statistical Bayesian T-test and generalized T-test tests, and combines it with the weight vector to get a new ranking score. However, the algorithm is time-consuming [22].

III. Proposed algorithm

A. Relevant filter feature selection algorithm

1) Information gain: Information gain is a measure of the dependence between features and class labels. Information gain is also one of the most popular Filter feature selection methods [23], due to its high computational efficiency and ease of interpretability. For each feature, the importance of the feature is measured primarily by how much information about the class is obtained from the feature. When selecting the optimal feature subset, we usually choose features that bring more information about the class. The degree of feature usefulness depends on the degree of entropy reduction of the class [24] when the corresponding feature is considered separately. The information gain (IG) of the feature X and the class Y is defined as:

$$IG(X,Y) = H(X) - H(X|Y) \tag{1}$$

Entropy is a measure of the uncertainty associated with a random variable. The information entropy $H(X)$ of the random variable X is defined as:

$$H(X) = -\sum_{i} P(x_i) * \log_2 P(x_i) \tag{2}$$

Where x_i represents a specific value of the random variable X, and $P(x_i)$ represents the probability that the variable X takes the value x_i.

The maximum value of the information gain is 1, and the higher the information gain value, the better the feature is. Generally, when the optimal subset is selected, the feature of the previous k information gain value is selected. In the classification problem, the random variable X represents the feature, and the random variable Y usually represents the class label.

2) Gini index: Gini index is a method to measure the ability of feature classification. Given the category C, the Gini index of characteristic f is defined as:

$$GiniIndex(f) = 1 - \sum_{i=1}^{C} [p(i|f)]^2 \tag{3}$$

For binary classification, the maximum Gini index is 0.5. The smaller the Gini index is, the better the feature is. Generally, when selecting the optimal subset, the top k features with the highest Gini indices are selected.

B. Wrapper feature selection algorithm CFGVM

The main idea of CFGVM is as follow: general vector machine (GVM) [25, 26, 27] is improved by binary ant lion optimizer (BALO) [28] to assign different weights to different class to proposed CGVM algorithm, then BALO is used as feature selection to choose the optimal features.

C. Proposed algorithm

The advantages of Filter algorithm are fast, scalable and independent of the classification algorithm. The disadvantage is that the classification $accuracy$ may not be very high [29]. The advantage of Wrapper algorithm is that the search process of feature subset and the process of model selection are interactive, and the dependency of feature is taken into account [30]. But its disadvantage is that it has high computational cost when constructing classifier. Because the Filter and Wrapper algorithm has their own advantages and disadvantages, we propose a hybrid Filter-Wrapper algorithm for high-dimensional small samples, which can eliminate most of the redundant and irrelevant features through filter algorithm, reduce the number of feature subsets on a large scale, and then

select the optimal feature subset through Wrapper algorithm to further improve the classification performance. In the Filter algorithm, we use the hybrid filter algorithm, mainly by combining the first n features of different filter algorithms to select better features. Wrapper algorithm chooses the algorithm CFGVM. The algorithm of UIG-CFGVM is: firstly, the top n features with the highest information gain indices and the top n features with the highest Gini indices selected by Filter, then the top n features of the two methods are combined to get the UIG algorithm, Furthermore, BALO algorithm improves GVM algorithm to select the optimal cost weights, and then BALO algorithm chooses the optimal features. The value of n is obtained by the experiment. The top m features are the union of the top n features obtained by the Info and the top n features obtained by the Gini.The proposed hybrid Filter-Wrapper algorithm 1 is as follows:

Algorithm 1 Hybrid Filter-Wrapper algorithm

input: n Number of features selected by a single Filter method; m Number of features of the data subset

output: $featuresub$ Optimal feature subset

1: Sort all features by the Info, select the top n features
2: Sort all features using the Gini method, select the top n features
3: The top m features are the union of the top n features obtained by the Info and the top n features obtained by the Gini
4: Select a subset of data based on the top m features
5: Execute Wrapper algorithm on data subset, the CFGVM is used as Wrapper algorithm
6: Get the best feature subset $featuresub$

IV. Experimental results and analysis

A. Datasets and experimental environment

To validate the performance of the proposed algorithm, seven biomedical high-dimensional imbalanced small sample datasets are used in the experiments performed. $ALLAML$, GLI_85 and $Leukemia$ from the scikit-feature selection repository database [31], $DCBCL$ from the Gene Expression Model Selector database [32]. Leu, MLL and $SRBCT$ datasets from literature [33]. Apart from SRBCT, other six datasets are two-class datasets. There are four categories in $SRBCT$ dataset which is convered into two categories. The names of the four categories are Burkitt's lymphom, rhabdomyosarcoma, the Ewing family of tumors and neuroblastoma. When converted to binary classification, Burkitt lymphoma is considered as one category and the other three categories as another. Table I shows the detailed information of the datasets. The final result is the average results of 20 times.

B. Evaluation metric and function

To compare the performance of different algorithms, we use the following evaluation indicators: $Accuracy$, $True\ positive\ rate$ (TPR), $False\ positive\ rate$ (FPR), $Area\ Under\ Curve$ (AUC), $F\text{-}measure$ and $G\text{-}mean$, f_n. f_n represents the number of the selected feature. The corresponding formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{8}$$

$$F\text{-}measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{9}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP + TN} * \frac{TN}{TN + FP}} \tag{10}$$

Where $True\ positive$ (TP) and $True\ negative$ (TN) represent the number of majority and minority class samples correctly classied, respectively. $False\ positive$ (FP) and $False\ negative$ (FN) represent the number of majority and minority class samples mistakenly clssified, respectively. TPR is the value of predicted minority class classified corretly. FPR is the value of predicted majority class mistakenly classified as minority class. $F\text{-}measure$ is weighted harmonic mean of $Precision$ and $Recall$. $Accuracy$ is the ratio of the number of correctly predicted samples to the number of all the predicted samples. $G\text{-}mean$ is a comprehensive indicator. AUC stands for the area under the ROC curve. Higher value of $Accuracy$, TPR , AUC, $F\text{-}measure$, indicate the better result, while smaller value of FPR indicates the better result.

C. Comparision of NBG algorithm and its similar algorithm

To better verify the performance of the proposed hybrid algorithm, this paper mainly compares with similar algorithms shown in table II. 20 features are selected after the Info and Gini algorithms, while the UIG is used to carry out the union operation by retaining 10 features with the Info and 10 features with the Gini. The complete dataset is split as: 80% of the data set is used for training and 20% for testing by random stratified sampling. In the Info-CFGVM, Gini-CFGVM and UIG-CFGVM, the scale of BALO is set to 15, and the number of iterations is set to 15. Because the number of features is an integer, we take the number of features f_n as an upward integer when they are used to count the number of features.

Table III shows the experimental results of various comparison algorithms. The best results for each indicator are expressed in bold black. Specific analysis as follows:

(1) Consider the results of different algorithms under the same Filter algorithm. When the Filter is Info, Info-CFGVM relative to the Info-GVM in seven data sets in terms of $Accuracy$, TPR, FPR, AUC, $G\text{-}mean$, $F\text{-}measure$ results are better. when the Filter is Gini, Gini-CFGVM relative to Gini-GVM in seven data sets on five classification indexes are better. When the UIG is used by Filter, the UIG-CFGVM and UIG-GVM on $DCBCL$, GLI_85, Leu with respect to $Accuracy$, TPR,

TABLE I
DATA DESCRIPTIONS USED IN THE EXPERIMENT

Name	Number of features	Sample size	Minority Sample size	Majority Sample size	Imbalance rate(IR)	Source
ALLAML	7129	72	25	47	1.88	ASU
DLBCL	5469	77	19	58	3.052	GEMS
GIL_85	22283	85	26	59	2.269	ASU
Leu	3571	72	25	47	1.88	NCBI
Leukemia	7070	72	25	47	1.88	ASU
MLL	5848	72	20	52	2.6	NCBI
SRBCT	2308	83	11	72	6.545	NCBI

TABLE II
DESCRIPTION OF CONTRAST ALGORITHMS

Algorithm	Detailed description
Info-GVM	Info is used as filter algorithm, GVM is used as classification algorithm
Gini-GVM	Gini is used as filter algorithm, GVM is used as classification algorithm
UIG-GVM	UIG is used as filter algorithm, GVM is used as classification algorithm
Info-CFGVM	Info is used as filter algorithm, then CFGVM is used as wrapper algorithm
Gini-CFGVM	Gini is used as filter algorithm, then CFGVM is used as wrapper algorithm
UIG-CFGVM	UIG is used as filter algorithm, then CFGVM is used as wrapper algorithm

TABLE III
EXPERIMENTAL RESULTS OF UIG-CFGVM AND OTHER SIMILAR ALGORITHMS ON THE TESTING DATASET

Dataset	Method	Accuracy	TPR	FPR	AUC	G-mean	F-measure	fn
ALLAML	Info-GVM	0.9333	**1**	0.1	0.95	0.9487	0.9091	20
	Gini-GVM	0.7333	0.8	0.3	0.75	0.7483	0.6667	20
	UIG-GVM	**1**	**1**	**0**	**1**	**1**	**1**	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	12
	Gini-CFGVM	0.8	**1**	0.3	0.85	0.8367	0.7692	6
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	**6**
DCBCL	Info-GVM	0.875	0.5	**0**	0.75	0.7071	0.6667	20
	Gini-GVM	0.875	0.5	**0**	0.75	0.7071	0.6667	20
	UIG-GVM	0.9375	0.75	**0**	0.875	0.866	0.8571	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	12
	Gini-CFGVM	0.875	0.75	0.0833	0.8333	0.8292	0.75	12
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	11
GLI_85	Info-GVM	0.8889	1	0.1667	0.9167	0.9129	0.8571	202
	Gini-GVM	0.7778	0.3333	**0**	0.6667	0.5774	0.5	20
	UIG-GVM	0.9444	1	0.0833	0.9583	0.9574	0.9231	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	9
	Gini-CFGVM	0.7778	1	0.3333	0.8333	0.8165	0.75	14
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	7
Leu	Info-GVM	0.9333	0.8	**0**	0.9	0.8944	0.8889	20
	Gini-GVM	0.5333	0.6	0.5	0.55	0.5477	0.4615	20
	UIG-GVM	0.9333	0.8	**0**	0.9	0.8944	0.8889	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	9
	Gini-CFGVM	0.8	0.8	0.2	0.8	0.8	0.7273	8
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	8
Leukemia	Info-GVM	0.9333	1	0.1	0.95	0.9487	0.9091	20
	Gini-GVM	0.6667	0.8	0.4	0.7	0.6928	0.6154	20
	UIG-GVM	**1**	**1**	**0**	**1**	**1**	**1**	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	12
	Gini-CFGVM	0.8	1	0.3	0.85	0.8367	0.7692	12
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	9
MLL	Info-GVM	0.6667	1	0.4545	0.7727	0.7385	0.6154	20
	Gini-GVM	0.7333	0	**0**	0.5	0	0	20
	UIG-GVM	**1**	**1**	**0**	**1**	**1**	**1**	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	11
	Gini-CFGVM	0.5333	1	0.6364	0.6818	0.603	0.5333	11
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	**7**
SRBCT	Info-GVM	0.9444	0.6667	**0**	0.8333	0.8165	0.8	20
	Gini-GVM	0.7222	0.6667	0.2667	0.7	0.6992	0.444	20
	UIG-GVM	**1**	**1**	**0**	**1**	**1**	**1**	20
	Info-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	7
	Gini-CFGVM	0.9444	0.6667	**0**	0.8333	0.8165	0.8	12
	UIG-CFGVM	**1**	**1**	**0**	**1**	**1**	**1**	**6**

FPR, AUC, $G\text{-}mean$ and $F\text{-}measure$ have good results, although from the $Accuracy$, TPR, FPR, AUC,$G\text{-}mean$ values in dataset $ALLAML$, $Leukemia$, MLL, $SRBCT$ are all 1, 1, 0, 1, 1, but the UIG-CFGVM needs fewer features than UIG-GVM, so its performance is better. This shows that CFGVM algorithm are better than single classification algorithm GVM in high-dimensional imbalanced small sample classification problem. It also shows that the performance of the proposed algorithm in this paper is the best in all algorithms. They can not only get the highest classification index, but also get the least number of features.

(2) Consider the results of different algorithms under the same Wrapper algorithm. In the case that the Wrapper algorithm is the same algorithm, using different Filter algorithm Info, Gini, UIG, the experimental results show that when the Filter is UIG , the classification performance is better than using the Filter is Info, Gini on 7 datasets. That is to say, UIG-GVM outperforms Info-GVM and Gini-GVM in terms of $Accuracy$, TPR, FPR, AUC, $G\text{-}mean$ and $F\text{-}measure$ on seven datasets. The imbalanced classfication performance of UIG-CFGVM is better than Info-CFGVM, Gini-CFGVM on $Accuracy$, TPR, FPR, AUC, $G\text{-}mean$ and $F\text{-}measure$ metrics on seven datasets. This shows that UIG combined with Wrapper algorithm has better classification performance than Info, Gini algorithm combined with Wrapper algorithm.

V. Conclusion

For the high-dimensional imbalanced small sample classification problem, due to the small number of high-dimensional samples and the imbalance of data, it brings great difficulty to the classification task and it takes a long time to complete the training of the model. In order to overcome these problems, data preprocessing is an effective algorithm. As an important preprocessing technology, feature selection can not only delete redundant features to improve classification performance, but also help identify key features related to classification problems. Because each feature selection algorithm has its own advantages and disadvantages, the performance of filter feature selection and Wrapper feature selection can be further improved by combining the advantages of filter feature selection and Wrapper feature selection. Based on this, this paper proposes two hybrid Filter-Wrapper algorithms UIG-CFGVM. The proposed algorithm first solve the high-dimensional problem by the Filter method, and then solve the imbalance problem and the over-fitting problem of the small sample through the Wrapper algorithm. The proposed algorithm is validated by experiments on seven different genetic data sets. The experimental results show that the proposed algorithm can improve the classification performance of high-dimensional imbalanced small sample classification problems. Compared with other seven similar algorithms, the proposed algorithm has better classification performance and requires fewer features.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China under Grant No. 61402210, The Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2018-k12, Ministry of Education - China Mobile Research Foundation under Grant No. MCM20170206, Major National Project of High Resolution Earth Observation System under Grant No. 30-Y20A34-9010-15/17, State Grid Corporation Science and Technology Project under Grant No. SGGSKY00FJJS1800403 and No.522722160071, Program for New Century Excellent Talents in University under Grant No. NCET-12-0250, and Strategic Priority Research Program of the Chinese Academy of Sciences with Grant No. XDA03030100.

References

[1] Hanaa Salem, Gamal Attiya, and Nawal El-Fishawy. Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing*, 50:124–134, 2017.

[2] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi. Hidden markov models for cancer classification using gene expression profiles. *Information Sciences*, 316(C):293–307, 2015.

[3] V. Boln-Canedo, N. Snchez-Maroo, A. Alonso-Betanzos, J. M. Bentez, and F. Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences An International Journal*, 282(5):111–135, 2014.

[4] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(12):245–271, 1997.

[5] N. Hoque, H. A. Ahmed, D. K. Bhattacharyya, and J. K. Kalita. A fuzzy mutual information-based feature selection method for classification. *Fuzzy Information & Engineering*, 8(3):355–384, 2016.

[6] Muhammad Summair Raza and Usman Qamar. An incremental dependency calculation technique for feature selection using rough sets. *Information Sciences*, 343-344:41–65, 2016.

[7] Shun Guo, Donghui Guo, Lifei Chen, and Qingshan Jiang. A centroid-based gene selection method for microarray data classification. *Journal of Theoretical Biology*, 400:32–41, 2016.

[8] Subhajit Kar, Kaushik Das Sharma, and Madhubanti Maitra. Gene selection from microarray gene expression data for classification of cancer subgroups employing pso and adaptive k-nearest neighborhood technique. *Expert Systems with Applications*, 42(1):612–627, 2015.

[9] Beatriz A Garro, Katya Rodríguez, and Roberto A Vázquez. Classification of dna microarrays using artificial neural networks and abc algorithm. *Applied Soft Computing*, 38:548–560, 2016.

[10] Hualong Yu, Guochang Gu, Haibo Liu, Jing Shen, and Jing Zhao. A modified ant colony optimization algorithm for tumor marker gene selection. *Genomics, proteomics & bioinformatics*, 7(4):200–208, 2009.

[11] Yanqiu Wang, Xiaowen Chen, Wei Jiang, Li Li, Wei Li, Lei Yang, Mingzhi Liao, Baofeng Lian, Yingli Lv, Shiyuan Wang, et al. Predicting human microrna precursors based on an optimized feature

subset generated by ga–svm. *Genomics*, 98(2):73–78, 2011.

[12] Ivana Vukusic, Sushma Nagaraja Grellscheid, and Thomas Wiehe. Applying genetic programming to the prediction of alternative mrna splice variants. *Genomics*, 89(4):471–479, 2007.

[13] A Sharma, S Imoto, and S Miyano. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 9(3):754–764, 2012.

[14] Seokho Kang, Dongil Kim, and Sungzoon Cho. Efficient feature selection-based on random forward search for virtual metrology modeling. *IEEE Transactions on Semiconductor Manufacturing*, PP(99):1–1, 2016.

[15] P Mohapatra, Sreejit Chakravarty, and PK Dash. Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm and Evolutionary Computation*, 28:144–160, 2016.

[16] HyunJi Kim, Byong Su Choi, and Moon Yul Huh. Booster in high dimensional data classification. *IEEE transactions on knowledge and data engineering*, 28(1):29–40, 2016.

[17] Kup-Sze Choi, Yugu Zeng, and Jing Qin. Using sequential floating forward selection algorithm to detect epileptic seizure in eeg signals. In *2012 IEEE 11th International Conference on Signal Processing*, volume 3, pages 1637–1640. IEEE, 2012.

[18] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

[19] Jadwiga R Bienkowska, Gul S Dalgin, Franak Batliwalla, Normand Allaire, Ronenn Roubenoff, Peter K Gregersen, and John P Carulli. Convergent random forest predictor: methodology for predicting drug response from genome-scale data applied to anti-tnf response. *Genomics*, 94(6):423–432, 2009.

[20] Md Eamin Rahman, Rashedul Islam, Shahidul Islam, Shakhinur Islam Mondal, and Md Ruhul Amin. Mirann: A reliable approach for improved classification of precursor microrna using artificial neural network model. *Genomics*, 99(4):189–194, 2012.

[21] Meihong Zhu and Song Jie. An embedded backward feature selection method for mclp classification algorithm. *Procedia Computer Science*, 17:1047–1054, 2013.

[22] Shruti Mishra and Debahuti Mishra. Svm-bt-rfe: An improved gene selection framework using bayesian t-test embedded in support vector machine (recursive feature elimination) algorithm. *Karbala International Journal of Modern Science*, 1(2):86–96, 2015.

[23] Thomas M Cover and Joy A Thomas. *Elements of information theory /*. 2003.

[24] Tan Feng, Xuezheng Fu, Yanqing Zhang, and Anu G. Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Computing*, 12(2):111–120, 2008.

[25] Hong Zhao. General vector machine. 2016.

[26] Qingguo Zhou, Fang Feng, Zebang Shen, Rui Zhou, Meng-Yen Hsieh, and Kuan-Ching Li. A novel approach for mobile malware classification and detection in android systems. *Multimedia Tools and Applications*, 78(3):3529–3552, 2019.

[27] Feng Fang, Qingguo Zhou, Zebang Shen, Xuhui Yang, Lihong Han, and Jin Qiang Wang. The application of a novel neural network in the detection of phishing websites. *Journal of Ambient Intelligence & Humanized Computing*, (13):1–15, 2018.

[28] E. Emary, Hossam M. Zawbaa, and Aboul Ella Hassanien. Binary ant lion approaches for feature selection. *Neurocomputing*, 213:54–65, 2016.

[29] Mary Walowe Mwadulo. A review on feature selection methods for classification tasks. *International Journal of Computer Applications Technology and Research*, 5(6):395–402, 2016.

[30] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[31] scikit-feature feature selection. *http://featureselection.asu.edu/datasets.php*.

[32] Gene expression model selector. *http://www.gems-system.org/*.

[33] Kun Yang, Zhipeng Cai, Jianzhong Li, and Guohui Lin. A stable gene selection in microarray data analysis. *Bmc Bioinformatics*, 7(1):228, 2006.