ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

GA-NobRFE-SVM: A New Algorithm for Classification of Unbalanced Gene Data

Botao Hu[1,a]

[1]*School of Information Engineering, Nanjing University of Finance & Economics, Nanjing, China*
[a]*1120201140@stu.nufe.edu.cn*

Abstract: *As an important branch of bioinformatics, gene microarray data analysis has become one of the important frontier fields in life sciences. Because of the high cost of microarray experiment, gene expression profile data shows the characteristics of small samples size, high dimensionality and category imbalance between samples. In this case, the traditional feature selection method is difficult to obtain good results. this paper proposes unbiased SVM-RFE (NobSVM-RFE). Compared with traditional feature selection algorithm, NobSVM-RFE algorithm can obtain better feature subset and reduce computation cost. Combining GASMOTE with NobSVM-RFE, this paper proposes a three-stage algorithm GA-NobRFE-SVM, which includes balancing algorithm, feature selection algorithm and classifier. The experimental results show that GA-NobRFE-SVM can effectively improve the classification performance of unbalanced gene data.*

Keywords: *Unbalanced gene data, Balancing algorithm, GASMOTE, Feature selection, NobSVM-RFE*

1. Introduction

In recent years, DNA microarray technology has greatly promoted the development of bioinformatics[1]. Microarray technology can quickly and accurately capture the expression of a large number of genes, thus providing a large number of unbalanced gene datasets for genomics, tumor diagnosis, pharmacogenomics and other fields, which also makes the classification task of unbalanced gene datasets become one of the important issues in bioinformatics. The classification task of unbalanced gene datasets is a new challenge in the field of machine learning because unbalanced gene datasets often have the characteristics of high dimensionality and small samples, and there is a serious problem of unbalanced distribution between samples[2].

In order to solve the problem of dataset imbalance, Chawla et al.[3] proposed the SMOTE algorithm. Furthermore, Kun J et al.[4] proposed the GASMOTE algorithm based on SMOTE, which uses genetic algorithm to search for the optimal sampling rate of minority class instances, and oversampling minority class instances at the optimal sampling rate.

In order to reduce the dimension of the dataset, Guyon et al.[5] proposed a feature selection algorithm using support vector machines, called Support Vector Machine Recursive Feature Elimination(SVM-RFE). The existing improved algorithms of SVM-RFE, such as SVM-T-RFE, SVM-BT-RFE[6], and SVM-PSO-RFE, filter features using weighted coefficients of biased SVM, and bias b affects the generalization performance and classification accuracy of SVM, Therefore, these improved algorithms filter out sub-optimal feature subsets.

To improve SVM-RFE performance, this paper proposes a new feature selection algorithm unbiased SVM-RFE (NobSVM-RFE). Then combined with the existing unbalanced data processing methods, this paper proposes a new three-stage algorithm GA-NobRFE-SVM, in which GASMOTE reduces the imbalance between samples, NobSVM-RFE algorithm filters the optimal feature subset, and SVM classifies the dataset. Experiments on three unbalanced gene datasets show that GA-NobRFE-SVM was significantly superior to other three-stage algorithms.

2. GASMOTE

In 2016, Kun J et al.[4] pointed out that existing SMOTE-based algorithms use the same sampling rate for all instances of minority classes, resulting in sub-optimal performance. Based on SMOTE, they propose the GASMOTE algorithm, which uses genetic algorithms to find optimal sampling rates for

ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

different minority class instances. The instances generated by the optimal sampling rate have higher quality and the distribution of minority class instances is more uniform. The specific steps as follows:

Step 1: Coding and initialization. Firstly, GASMOTE will generate a population with size P. Then, GASMOTE uses a single individual X^j in a population to represent the combination of sampling rates for all instances, as shown in formula (1):

$$X^j = \left(N_1^j, N_2^j, ..., N_M^j\right),\ j=1, 2, \ldots, P \tag{1}$$

M represents the number of instances of minority cases, P represents population size, and N represents the sampling rate of minority instances.

Step 2: Selection. The population is arranged in descending order according to the fitness function value f and the lower ranked individuals are deleted. Fitness function f is shown in formula (2):

$$f = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{2}$$

Where TP is true positive, FP is false positive, TN is true negative and FN is false negative.

Step 3: Crossover.

Step 4: Mutation.

Step 5: check the termination condition. The termination condition is that if the number of iterations is greater than the threshold, the algorithm outputs the instances generated by the optimal sampling rate; otherwise, return to step 2

3. Proposed algorithm

3.1 Unbiased SVM-RFE algorithm

Poggio[7] pointed out that the SVM optimization problem for positive definite kernel functions does not require bias b. Reference [8] indicate that SVM with Gaussian Kernel do not need bias b, so we propose a new feature selection algorithm unbiased SVM-RFE (NobSVM-RFE). Before presenting the NobSVM-RFE algorithm, introduce the unbiased SVM.

Given a dataset $\{x_a, y_a\}_{a=1}^M$ with M instances,where $x_a \in R^N$, $y_a \in R$. Because there is no bias b constraint, the optimization problem of unbiased SVM is shown in Formula (3):

$$\min \frac{1}{2}\|\omega\|^2 + C\sum_{a=1}^{M}\xi_a$$

$$s.t. y_a < \omega \cdot x_a > \geq 1-\xi_a,\ \xi_a \geq 0, a = 1,...,M \tag{3}$$

With the KKT condition, the objective function of formula (3) can be rewritten to formula (4):

$$L=\frac{1}{2}\|\omega\|^2 + C\sum_{a=1}^{M}\xi_i - \sum_{a=1}^{M}\alpha_a(y_a < \omega \cdot x_a > -1+\xi_a) - \sum_{a=1}^{M}\beta\xi_a \tag{4}$$

Where α_a and β_a are non negative Lagrange multipliers.

Finding the Partial Derivative of L with respect to α_a and β_a and simplify them to obtain the dual optimization problem of unbiased SVM, as shown in formula (5).

$$\min \frac{1}{2}\sum_{a=1}^{M}\sum_{b=1}^{M} y_a y_b \alpha_a \alpha_b < x_a, x_b > - \sum_{a=1}^{M}\alpha_a$$

$$s.t.\ 0 \leq \alpha_a \leq C, a = 1,...M \tag{5}$$

Reference [8] points out that the computation, generalization performance and classification accuracy of unbiased SVM are superior to those of ordinary support vector machine. Therefore, if unbiased SVM is used to train the original dataset during recursive elimination, NobSVM-RFE will

ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

filter out a better feature subset.

The specific steps of the NobSVM-RFE algorithm are as follows:

Step 1: Initialize the original dataset S = [1,2,... n], the maximum number of features eliminated per iteration max_s, the number of features in the feature subset max_n (max_n < = | S |), cost parameter C, and the Gaussian kernel parameter gamma. Set feature subset R as empty set.

Step 2: Repeat 3-7 until | R | = n-max_s.

Step 3: On the dataset S, train an unbiased SVM with Gaussian kernel, where the cost parameter of the unbiased SVM is C and the kernel parameter is gamma.

Step 4: Calculate the weight vector w of unbiased SVM.

Step 5: Calculate the score for each feature in S.

Step 6: Repeat 7, min (max, n-max_n-| R |) times.

Step 7: Search for the feature e with the lowest score in S, set R to [e, R], and set S to S – {e}

Finally, output feature subset S.

3.2 GA-NobRFE-SVM

Combining GASMOTE, NobSVM-RFE and SVM, this paper proposes a new three-stage algorithm GA-NobRFE-SVM. The specific steps are as follows:

Step 1: Call the algorithm GASMOTE to balance the original dataset Pre_data and get the balanced dataset BalData.

Step 2: Call algorithm NobSVM-RFE to get the optimal feature subset RmData.

Step 3: Use RmData as the training set, train an SVM and calculate the SVM classification accuracy.

The flowchart of GA-NobRFE-SVM is shown in Figure 1:

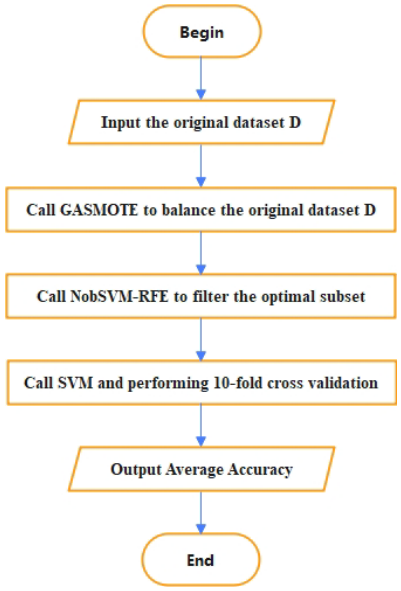

Figure 1: GA-NobRFE-SVM flow chart.

4. Experiments and result analysis

4.1 Dataset Introduction

In the experiments, three datasets are used, namely Armstrong-2002-v1, Golub-1999-v1 and Pomeroy-2002-v1. The specific parameters of the datasets is shown in Table 1:

ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

Table 1: Dataset Information.

Dataset	F_number	S_number	S_distribution
Armstrong-2002-v1	12582	72	1:2
Golub-1999-v1	7129	72	47:25
Pomeroy_2002-v1	7129	34	25:9

Where F_number is the number of features, S_number is the number of samples, S_distribution is the degree of imbalance.

4.2 Experimental parameters setting

In the experiments, the balancing algorithms all use the smote_variants toolbox, feature selection algorithms use the yan_prtools toolbox, and the classifier SVM uses the Libsvm toolbox.

For the algorithm GA NobRFE-SVM, the value range of population quantity P is [2, 4, 6, 8, 10], the lowest sampling rate L is set to 0, and the highest sampling rate H is set to 1，the value range of the cost parameter C is [0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 500, 1 000] , the value range of the kernel parameter gamma is [0.001, 0.01, 0.1, 0.2, 0.4, 0.8, 1, 2, 5, 10, 20, 50, 100]。

4.3 Experimental results of Armstrong-2002-v1

In the experiment, each dataset will be subject to two groups of experiments, namely three-stage integration experiment and validation experiment. The three-stage integration experiment used to screen the best three-stage algorithm. The validation experiment mainly tests and judges the performance of different algorithms on multiple number of genes.

Table 2: Three-stage integration experiment table of Armstrong-2002-v1.

Accuracy	CIFE	MRMR	LinearSVM-RFE	NobSVM-RFE	SVM-BT-RFE
CURESMOTE	**85.965(+/-1.129)**	**91.228(+/-2.962)**	**94.737(+/-1.886)**	**98.246(+/-0.962)**	**87.719(+/-2.849)**
GASMOTE	**97.619(+/-1.367)**	**96.667(+/-1.428)**	**95.699(+/-1.047)**	**98.851(+/-0.251)**	**97.701(+/-0.251)**
GaussianSMOTE	**68.421(+/-1.886)**	**85.965(+/-2.924)**	**68.421(+/-1.886)**	**70.175(+/-2.924)**	**94.737(+/-1.147)**
SMOTETomek	**91.228(+/-1.924)**	**89.474(+/-1.595)**	**89.474(+/-1.595)**	**92.982(+/-2.962)**	**96.246(+/-1.041)**
SVMSMOTE	**82.353(+/-2.415)**	**97.125(+/-1.637)**	**98.246(+/-0.262)**	**90.741(+/-0.858)**	**87.719(+/-1.891)**

Table 3: Armstrong-2002-v1 validation experiment table.

Accuracy	20	40	60	80	100
CURESMOTE+ NobSVM-RFE	**94.737(+/-0.633)**	**94.737(+/-1.952)**	**100.0(+/-0.0)**	**84.211(+/-1.147)**	**98.246(+/-0.962)**
GA-NobRFE-SVM	**96.552(+/-2.236)**	**100.0(+/-0.0)**	**100.0(+/-0.0)**	**100.0(+/-0.0)**	**98.851(+/-0.251)**
GaussianSMOTE+ SVM-BT-RFE	**100.0(+/-0.0)**	**96.552(+/-1.427)**	**100.0(+/-0.0)**	**84.211(+/-1.125)**	**94.737(+/-1.147)**
SMOTETomek+ SVM-BT-RFE	**100.0(+/-0.0)**	**100.0(+/-0.0)**	**94.737(+/-3.251)**	**96.552(+/-2.142)**	**96.246(+/-1.041)**
SVMSMOTE+ MRMR	**89.474(+/-0.514)**	**100.0(+/-0.0)**	**89.474(+/-1.375)**	**100.0(+/-0.0)**	**97.125(+/-1.637)**

As shown in Table 2 and Table 3, although the classification accuracy of GA-NobRFE-SVM algorithm is slightly lower than SMOTETomek+SVM-BT-RFE in the Armstrong-2002-v1 dataset with 20 pre-selected genes, when the number of pre-selected genes reaches 40 or more(40,60,80,100), GA-NobRFE-SVM has the highest classification accuracy.

ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

4.4 Experimental results of Pomeroy-2002-v1

Table 4: Three-stage integration experiment table of Pomeroy-2002-v1.

Accuracy	CIFE	MRMR	LinearSVM-RFE	NobSVM-RFE	SVM-BT-RFE
CURESMOTE	90.0(+/- 1.649)	86.0(+/- 1.667)	90.0(+/- 0.649)	92.0(+/- 1.967)	88.0(+/- 1.333)
GASMOTE	95.0(+/- 1.333)	98.462(+/- 1.154)	97.143(+/- 2.999)	100.0(+/- 0.0)	98.462(+/- 2.154)
GaussianSMOTE	65.0(+/- 1.428)	75.0(+/- 2.096)	87.5(+/- 1.428)	65.0(+/- 2.428)	92.0(+/- 1.596)
SMOTETomek	94.0(+/- 0.125)	98.0(+/-1.667)	92.0(+/- 1.967)	90.0(+/- 1.909)	94.0(+/- 1.637)
SVMSMOTE	86.0(+/- 1.396)	88.0(+/- 1.596)	100.0(+/- 0.0)	93.333(+/- 0.667)	98.0(+/- 1.625)

Table 5: Pomeroy-2002-v1 vaildation experiment table.

Accuracy	20	40	60	80	100
CURESMOTE+ NobSVM-RFE	86.0(+/- 2.007)	96.0(+/- 1.798)	92.0(+/- 1.967)	94.0(+/- 0.798)	92.0(+/- 1.967)
GA-NobRFE-SVM	96.0(+/- 2.128)	98.0(+/-1.667)	98.0(+/- 1.333)	98.2(+/- 1.112)	100.0(+/- 0.0)
GaussianSMOTE+ SVM-BT-RFE	94.0(+/- 1.798)	98.0(+/-0.125)	92.0(+/- 1.967)	92.0(+/- 2.659)	92.0(+/- 1.596)
SMOTETomek+ MRMR	88.0(+/- 2.967)	94.0(+/- 1.798)	96.0(+/- 1.798)	92.0(+/- 1.967)	98.0(+/- 1.677)
SVMSMOTE+ LinearSVM-RFE	96.0(+/- 2.128)	96.0(+/- 1.798)	98.0(+/-1.125)	96.0(+/- 2.125)	100.0(+/- 0.0)

As shown in Table 4 and Table 5, GA-NobRFE-SVM algorithm achieves optimal classification performance in the Pomeroy-2002-v1 dataset.

4.5 Experimental results of Golub-1999-v1

Table 6: Three-stage integration experiment table of Golub-1999-v1.

Accuracy	CIFE	MRMR	LinearSVM-RFE	NobSVM-RFE	SVM-BT-RFE
CURESMOTE	89.474(+/- 0.614)	88.421(+/- 1.470)	82.105(+/- 1.735)	98.296(+/- 1.476)	95.192(+/- 1.421)
GASMOTE	94.615(+/- 0.154)	90.667(+/- 0.844)	92.857(+/- 2.825)	100.0(+/- 0.0)	96.0(+/- 2.764)
GaussianSMOTE	73.684(+/- 0.449)	76.842(+/- 0.295)	67.368(+/- 1.098)	88.421(+/- 2.276)	94.737(+/- 1.125)
SMOTETomek	87.368(+/- 1.279)	89.474(+/- 1.531)	87.368(+/- 1.754)	82.105(+/- 1.735)	91.579(+/- 1.279)
SVMSMOTE	90.526(+/- 1.877)	78.947(+/- 1.315)	81.053(+/- 1.754)	87.368(+/- 1.632)	93.75(+/- 1.180)

Table 7: Golub-1999-v1 vaildation experiment table.

Accuracy	20	40	60	80	100
CURESMOTE+ NobSVM-RFE	100.0(+/- 0.0)	93.150(+/- 1.428)	94.737(+/- 1.333)	94.737(+/- 0.856)	98.296(+/- 1.476)
GA-NobRFE-SVM	92.593(+/- 0.252)	95.855(+/- 1.625)	100.0(+/- 0.0)	100.0(+/- 0.0)	100.0(+/- 0.0)
GaussianSMOTE+ SVM-BT-RFE	94.737(+/- 1.134)	92.198(+/- 1.566)	100.0(+/- 0.0)	100.0(+/- 0.0)	94.737(+/- 1.125)
SMOTETomek+ SVM-BT-RFE	95.755(+/- 1.677)	91.160(+/- 1.425)	89.474(+/- 1.252)	100.0(+/- 0.0)	91.579(+/- 1.279)
SVMSMOTE+ SVM-BT-RFE	96.198(+/- 1.320)	100.0(+/- 0.0)	100.0(+/- 0.0)	93.75(+/- 1.428)	93.75(+/- 1.180)

ISSN 2616-5775 Vol. 6, Issue 2: 10-15, DOI: 10.25236/AJCIS.2023.060202

As shown in Table 6 and Table 7, in the Golub-1999-v1 dataset, the classification accuracy of GA NobRFE-SVM gradually converges, achieving the best classification performance when the number of pre-selected genes reached 60 or more(60,80,100).

5. Conclusions

Firstly, aiming at the disadvantage of low performance of SVM-RFE, this paper proposes unbiased NobSVM-RFE algorithm. Combining GASMOTE with NobSVM-RFE algorithm, this paper proposes GA-NobRFE-SVM algorithm for classification task of unbalanced gene dataset. The experimental results show that the GA-NobRFE-SVM algorithm has the highest classification accuracy when the number of pre-selected genes reaches 60 or more(60,80,100). The theoretical analysis and experimental results show that GA-NobRFE-SVM can effectively handle the classification task of unbalanced gene datasets.

This paper only verified the effectiveness of GA-NobRFE-SVM in the binary unbalanced gene data set. The effectiveness of GA-NobRFE-SVM in the multi-class dataset still needs further testing.

References

[1] Bhadra T, Mallik S, Hasan N, et al. Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer[J]. BMC bioinformatics, 2022, 23(3): 1-19.
[2] Bommert A, Welchowski T, Schmid M, et al. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data[J]. Briefings in Bioinformatics, 2022, 23(1):1-13.
[3] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16(1): 321-357.
[4] Jiang K, Lu J, Xia K. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE[J]. Arabian journal for science and engineering, 2016, 41(8): 3255-3266.
[5] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1): 389-422.
[6] Mishra S, Mishra D. SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm[J]. Karbala International Journal of Modern Science, 2015, 1(2): 86-96.
[7] Poggio T, Mukherjee S, Rifkin R, et al. Technical Report AI Memo. No. 2001-011[J]. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA, 2001.
[8] Xiaojian Ding, Yinliang Zhao. Influence of bias b on generalization performance of support vector machine classification problems [J]. Journal of Automation, 2011, 37(9): 1105-1113.