Evaluating Ensemble Learning Impact on Gene Selection for Automated Cancer Diagnosis*

Ke Yan[1][0000−0002−1611−6636] and Huijuan Lu[2]

[1] College of Information Engineering, China Jiliang University, 310018 Hangzhou, China. yanke@cjlu.edu.cn
[2] College of Information Engineering, China Jiliang University, 310018 Hangzhou, China. hjlu@cjlu.edu.cn

Abstract. Modern artificial intelligence (AI) enabled research shows that cancers can be detected and diagnosed by classification of DNA micro-arrays in molecular level. DNA micro-arrays data has the special property of high-dimension with redundancy and noises that consists of t-housands of features. In this study, a novel hybrid feature selection framework is proposed based on ensemble learning to select the most important genes. Experimental results show that the proposed method effectively increases the classification accuracy compared to existing methods.

Keywords: Feature Selection · DNA Micro-array · ReliefF · Mutual Information Maximization · Ensemble Learning.

1 Introduction

Feature selection of DNA micro-arrays, followed by classification, is well recognized as a next generation technology for cancer diagnosis, prognosis and prediction [1]. The supervised classification process makes the computerized automatic diagnosis of various tumors possible.

We propose a novel extended GA (EGA) based hybrid feature selection framework to select important genes from expression data [2]. An ensemble machine learning structure is built to select important genes based on majority voting scheme.

2 Methodology

A hybrid feature selection framework is proposed to combine the filter based methods and wrapper based methods. The filter based methods include mutual information maximization and reliefF. Three classifiers, including CS-D-ELM [3], SVM and RoF [4], are combined with GA to select the important genes. In

* Supported by National Natural Science Foundation of China (grant numbers: 61850410531 and 61602431) and Zhejiang Provincial Natural Science Foundation of China (Nos. LY19F020016 and 2017C34003).

each GA process, new generation of feature subset is generated by crossover and mutation operations. The final selected feature subset is evaluated by a majority voting scheme between the three EGA algorithms. The overall flowchart the hybrid feature selection framework can be depicted as in Figure 1.

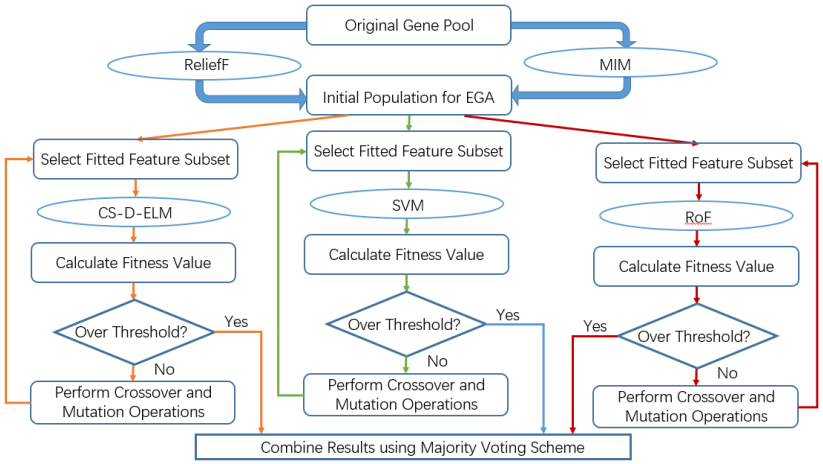

Fig. 1. The flowchart of the proposed hybrid gene selection framework.

3 Results

Four different cancer gene expression datasets were utilized for verification purposed, which include breast, lung, colon and leukemia. The number of samples, features and label distribution situations are listed in Table 1:

Table 1. The detailed information about three cancer diagnosis datasets.

Datasets	**# Sample**	**# Genes**	**Labels (# sample)**
Breast	19	24482	Non-relapse (7) / Relapse (12)
Lung	149	12535	Negtive (134) / Positive (15)
Colon	62	2000	Negtive (33) / Positive (33)
Leukemia	34	7130	ALL (20) / AML (14)

We compare the classification accuracy rates based on our method with three existing feature selection approaches: reliefF, MIM, MIM-GA [5]. The extreme learning machine (ELM) is selected to be the base classifier for fair comparison.

We force all four feature selection approaches to select the same number of features for the feature subsets. Ten different numbers of the feature subsets are designed and listed in Table 2.

Table 2. Ten different numbers of features for the selected feature subsets.

Datasets	**Number of Genes**									
Breast	6	18	32	56	88	112	144	156	168	196
Lung	4	32	73	96	114	128	144	156	186	202
Colon	19	38	64	96	114	126	158	178	198	216
Leukemia	7	48	80	96	124	150	168	178	188	198

The classification accuracy rates of different datasets are listed in Tables Table 3 to Table 6. It is noted that for each accuracy rate, 30 times repeated tests are performed to guarantee the generalization.

Table 3. Classification accuracy rates for the Breast dataset.

Methods	**Classification accuracy rates %**									
Proposed	84.95	87.21	92.30	94.71	96.26	97.12	95.28	94.95	95.38	96.82
ReliefF	73.68	68.42	73.68	73.68	78.94	78.94	73.68	73.68	78.94	78.94
MIM	78.27	66.38	72.41	74.38	76.81	78.37	77.46	80.92	79.38	77.90
MIM-GA	83.17	84.98	86.28	90.26	93.28	96.36	92.28	89.36	92.36	94.27

Table 4. Classification accuracy rates for the Lung dataset.

Methods	**Classification accuracy rates %**									
Proposed	93.28	90.28	92.99	94.93	96.28	98.47	96.82	96.28	97.73	98.05
ReliefF	74.28	63.38	66.86	70.38	71.38	73.47	76.28	75.10	73.28	75.86
MIM	80.92	74.28	76.82	79.38	81.46	84.29	83.42	82.04	83.92	84.28
MIM-GA	94.80	91.18	92.36	94.15	94.91	97.36	95.92	93.38	94.40	96.14

4 Conclusion

In this study, we introduced a hybrid feature selection method that combines filter based methods with wrapper based method. A sophisticated ensemble feature

Table 5. Classification accuracy rates for the Colon dataset.

Methods	**Classification accuracy rates %**									
Proposed	95.00	83.27	86.43	89.37	93.64	98.28	95.73	97.38	98.27	96.60
ReliefF	70.39	65.28	67.84	70.93	75.36	79.95	81.38	77.38	79.55	80.52
MIM	63.31	60.29	62.28	64.48	65.49	68.48	65.59	62.95	64.64	67.84
MIM-GA	83.40	77.63	81.28	83.01	85.87	89.14	93.37	89.98	91.47	92.75

Table 6. Classification accuracy rates for the Leukemia dataset.

Methods	**Classification accuracy rates %**									
Proposed	97.22	96.48	97.58	95.29	97.45	99.48	98.84	96.28	97.83	98.72
ReliefF	67.64	70.59	73.53	76.47	79.41	82.35	86.29	80.24	82.24	84.18
MIM	76.38	72.31	76.29	79.82	83.84	87.82	83.28	79.49	82.49	84.01
MIM-GA	97.50	94.48	95.30	96.28	97.39	98.02	94.29	95.30	95.54	97.24

selection framework is introduced to increase the generalization of GA. Experimental results show that the proposed method is suitable to handle various cancer diagnostic datasets, and provides highest classification accuracy among all compared methods.

References

[1] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530, 2002.

[2] Shutao Li, Xixian Wu, and Mingkui Tan. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing*, 12(11):1039–1048, 2008.

[3] Yanqiu Liu, Huijuan Lu, Ke Yan, Haixia Xia, and Chunlin An. Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification. *Computational intelligence and neuroscience*, 2016, 2016.

[4] Huijuan Lu, Lei Yang, Ke Yan, Yu Xue, and Zhigang Gao. A cost-sensitive rotation forest algorithm for gene expression data classification. *Neurocomputing*, 228:270–276, 2017.

[5] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256:56–62, 2017.