Stable and Accurate Feature Selection from Microarray Data with Ensembled Fast Correlation Based Filter

1st Aiguo Wang
School of Electronic Information Engineering
Foshan University
Foshan, China
wangaiguo2546@163.com

2nd Huancheng Liu
School of Electronic Information Engineering
Foshan University
Foshan, China
liuhuancheng83@163.com

3rd Jinjun Liu
School of Computer and Information Engineering
Chuzhou University
Chuzhou, China
jinjunl@chzu.edu.cn

4th Huitong Ding
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
ding_huitong@163.com

5th Jing Yang
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
jsjyj0801@163.com

6th Guilin Chen
School of Computer and Information Engineering
Chuzhou University
Chuzhou, China
glchen@chzu.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313533

Abstract—**Feature selection has been playing an important role in analyzing the high-dimension and low-sample-size gene expression profiles towards high classification performance of diseases and deep understanding of the underlying biological mechanisms. Besides classification performance, the stability of selected features is another non-ignorable factor in evaluating a feature selector, since stable feature selection results enhance the confidence of selected features for true biomarker discovery and further biological validation. In this study, we propose a novel feature selection method under the ensemble learning framework. Specifically, we take Fast Correlation Based Filter as the base feature selector to analyze subsamples of microarray data. We then present several aggregation methods to combine multiple feature subsets. Finally, two stability measures are used to quantify the robustness of feature selectors to data variations. Our comparative empirical study on publicly available datasets demonstrates the superiority of the proposed methods over its competitors in obtaining high stability scores and classification accuracy.**

Keywords—microarray data, feature selection, ensemble learning, stability, classification

I. Introduction

Microarray techniques enable researchers to measure the expression profiles of thousands of genes simultaneously and provide a convenient and objective way to diagnose cancers, discriminate tumor subtypes, predict survival of patients, and identify diseasing genes at the molecular level [1]. However, the inherent nature of high-dimensional and low-sample-size microarray data poses a serious challenge to effective mining and analyses. One commonly used and proven powerful way is to perform feature selection.

Feature selection, also called gene selection in the context of Genomics, aims to obtain the best predictive accuracy by removing irrelevant and redundant features from the original feature space. Accordingly, a wealth of feature selection methods, often grouped into filter, wrapper and embedded models, have been proposed and validated in various domains [2]. In microarray data analysis, besides the benefits of high classification accuracy, improved classifier training time, and reduced data collection cost, a good feature selection algorithm should provide insights into knowledge discovery and a small subset of genes best uncovering the underlying biological mechanisms, which, to a certain extent, requires it to have good stability. Stability means that a feature selector is robust to the perturbation of training data and the parameter values of learning algorithms. That is, a feature selection algorithm outputs similar features under different conditions [3]. In contrast, instability destroys the confidence of domain experts in identifying true biomarkers and performing further biological validation. Hence, high classification performance and stability are equally important in feature selection.

The sources of feature selection instability mainly include *biological mechanism level* (e.g., there exist multiple sets of true biomarkers), *data level* (high-dimensional low-sample-size data), and also *algorithm level* (e.g., an algorithm fails to consider stability), where the data level is a great source of instability. Accordingly, ensemble learning, using multiple accurate and diverse models, can be a solution [4]. To this end, we here propose an ensemble feature selection framework. Specifically, to better capture the non-linear relationships among features and relieve users from determining how many features to use, we take the filter, termed Fast Correlation Based Filter (FCBF), as the base selector. We then introduce five aggregators to combine multiple feature subsets into a final set and return it. Particularly, the main contributions of this study are as follows. (1) We present an ensembled FCBF based feature selector, and detail its components and introduce several aggregation methods. (2) We conduct experiments on public datasets in comparison with other five commonly used feature selectors in terms of classification performance and two stability measures. Results demonstrate its superiority in obtaining high stability scores and classification accuracy.

II. The Proposed Method

A. Ensemble Feature Selection Framework

According to the used base feature selectors, we categorize existing ensemble feature selectors into *homogeneous model* and *heterogeneous model*, where the former uses the same base selector and the later exploits different base selectors. We herein present our feature selector under the homogeneous

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2020A1515011499), the Major Special Projects of Anhui Province (Grant No. 201903A06020026), and the Anhui Provincial Natural Science Foundation (Grant No. 1908085MF211).

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on January 17,2021 at 08:37:26 UTC from IEEE Xplore. Restrictions apply.

framework, as shown in Fig. 1. In this scheme, the same base feature selector is trained with different subsets of the training data and one feature subset is returned for each subsample. This largely guarantees the diversity of training sets and enables us to parallelize the task. There are numerous way to obtain subsamples from the training data, such as k-fold cross validation, Bootstrap, and Gibbs sampling. Afterwards, an aggregation method combines multiple feature subsets into a final subset. Obviously, the choices of feature selectors and aggregators are two key components that largely determine the performance of an ensemble feature selector.

B. *Fast Correlation Based Filter*

As for the feature selector component, it outputs a subset or an ordered ranking of the original features. If the latter is used, we need a threshold step to get a subset of relevant but less redundant features. To better capture the highly non-linear relationships among features and relieve users from deciding how many features to keep, we take FCBF as the base selector.

Rather than evaluate each feature independently, FCBF measures the correlation between two features and uses the approximate Markov blanket technique to remove redundant features [5]. Specifically, FCBF calculates the C-correlation (correlation between a feature and the class) and F-correlation (correlation between two features), then filters out features whose C-correlation is less than a predefined threshold and removes redundant features with the identified approximate Markov blanket. Finally, FCBF returns an optimal subset.

C. *Aggregation Strategy*

Aggregator is another key component of an ensemble feature selector. Given k feature subsets $S_F = \{S_1, S_2, \ldots, S_k\}$ returned by k base feature selectors, we adopt the following strategies to get a subset S based on the frequency p_f of a feature f out of k. Suppose the union of the k feature subsets is SS,

1) Intersection Scheme: S is the intersection of subsets $\{S_1, S_2, \ldots, S_k\}$.

$$S = \{f | f \in SS, p_f = 1\} = \cap_{i=1}^{k} S_i \qquad (1)$$

2) Half Scheme: The criteria to include a feature f is that its frequency is no less than fifty percent.

$$S = \{f | f \in SS, p_f \geq 0.5\} \qquad (2)$$

3) Quarter Scheme: The criteria to select a feature f is that its frequency is no less than twenty-five percent.

$$S = \{f | f \in SS, p_f \geq 0.25\} \qquad (3)$$

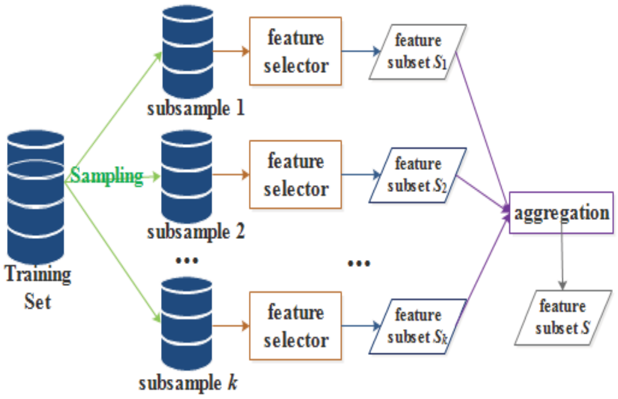


Fig. 1. Flowchart of the ensemble feature selection.

4) Union Scheme: S is the union of subsets $\{S_1, S_2, \ldots, S_k\}$.

$$S = \{f | f \in SS, p_f > 0\} = \cup_{i=1}^{k} S_i = SS \quad (4)$$

5) Union followed by FCBF Scheme: The union operation is first conducted and FCBF is then applied on the union.

III. Experimental Results and Analysis

A. *Experimental Setup*

To evaluate the proposed methods, we conduct extensive comparative experiments on four public microarray datasets, covering both binary and multi-classes cases [6], as shown in Table I. The last column "#SGR" means the ratio between the number of samples and the number of genes. As for feature selection algorithms, besides FCBF, we also include ReliefF, Mutual Information Maximization (MIM), Min-Redundancy Max-Relevance (MRMR), and Joint Mutual Information (JMI) as a comparison [7]. Since ReliefFF, MIM, MRMR, and JMI are ranking-based feature selectors, we experimentally choose the top twenty-five genes to obtain the final feature subset. For the ensemble feature selector, we take FCBF as the base feature selector with a five-fold sampling scheme and use different aggregators under the framework of Fig. 1.

For performance evaluation, we exploit stability measures to quantify the robustness of a feature selector and use the support vector machine models that are trained on the training sets and tested on the test set to evaluate the quality of selected features. We also report the classification performance of the case without using feature selection. Specifically, we use the ten-fold cross validation to split the microarray dataset into training sets and test sets and feature selection is performed on the training set towards an unbiased evaluation. Averaged accuracy and F1 of the ten results are reported. For the stability measures, we use the adjusted similarity (sim_L) and relative weighted consistency (CW_{rel}) to measure the robustness of a feature selector to the variation of training set in returning the same features [8]. Particularly, sim_L reports the average pairwise similarities between subsets of S_F and CW_{rel} uses the global frequency of each feature in SS to calculate the stability scores.

B. *Classification Performance*

Table II gives the classification performance of ensembled FCBF and their competitors on the experimental datasets. The column "Original" indicates the results without using feature selection, and FCBF_H, FCBF_Q, FCBF_U, and FCBF_F correspond to the four aggregation strategies given in *2*), *3*), *4*), and *5*), respectively. Particularly, the use of intersection scheme returns an empty S in some cases, and we omit it in the study. The best results for each dataset are shown in bold. From Table II, we observe that FCBF generally outperforms ReliefF, MIM, MRMR, and JMI in terms of accuracy and F1, which indicates the power of FCBF in selecting discriminant genes. We also observe that the ensembled FCBF obtains comparable and often better performance than FCBF and that FCBF_Q gets the best results, which shows their power.

TABLE I. Experimental Datasets

ID	Dataset	#Classes	#Samples	#Genes	#SGR
1	*COLON*	2	62 (40/22)	2000	0.031
2	*DLBCL*	2	77 (58/19)	7129	0.011
3	*LEUMEMIA*	3	72 (38/9/25)	5327	0.014
4	*SRBCT*	4	83 (29/25/11/18)	2308	0.036

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on January 17,2021 at 08:37:26 UTC from IEEE Xplore. Restrictions apply.

TABLE II. ACCURACY AND F1 COMPARISONS OF THE PROPOSED METHODS AND ITS COMPETITORS

Dataset	Original		ReliefF		MIM		MRMR		JMI		FCBF		FCBF_H		FCBF_Q		FCBF_U		FCBF_F	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
1	80.65	79.36	77.42	74.39	75.81	72.89	82.26	81.37	74.19	71.29	77.42	75.34	79.03	76.54	**83.87**	**82.39**	79.03	77.35	82.26	80.20
2	96.10	94.70	96.10	94.88	93.51	91.44	94.81	93.01	92.21	89.52	96.10	94.70	90.91	88.59	**98.70**	**98.25**	96.10	94.70	97.40	96.50
3	93.06	93.07	91.67	91.46	87.50	87.82	91.67	91.80	94.44	94.39	97.22	97.14	97.22	97.06	**98.61**	**98.59**	97.22	97.14	97.22	97.14
4	100.0	100.0	97.59	98.14	100.0	100.0	100.0	100.0	100.0	100.0	98.80	99.00	100.0	100.0	**100.0**	**100.0**	100.0	100.0	100.0	100.0

C. Comparison of Stability

For the purpose of this study, we here only present the stability comparisons of FCBF and its ensemble versions on the experimental datasets, as shown in Fig. 2. For each dataset, we use the adjusted similarity (sim_L) and relative weighted consistency (CW_{rel}) as stability measures. From Fig. 2, we observe that the ensembled FCBF outperforms FCBF in the majority of cases and that FCBF_Q tends to achieve a higher stability score. This verifies that the proposed methods have better stability.

Besides, we present the (average ± standard deviation) number of selected features of FCBF and its ensemble versions in Table III. According to the results of FCBF_H, FCBF_Q, FCBF_U, we observe that the number of selected features tends to increase with the decrease of inclusion criteria values. We also observe that the number of features returned by FCBF_Q is comparable to that of FCBF. Overall, according to the above results and analysis, FCBF_Q remains a priority in returning stable and discriminant features.

IV. CONCLUSION

As for the analysis of microarray data, high classification performance and stability are equally important in evaluating an feature selector, where high accuracy helps better classify cancers and tumor subtypes and stable feature selection results enhance the confidence of domain experts in further analysis of the identified biomarkers. In this study, we propose an ensemble-based feature selection framework that takes as the base selector the filter method Fast Correlation Based Filter. To combine multiple feature subsets into a final subset, we introduce five frequency-based aggregation methods. Besides, two stability measures (i.e., the adjusted similarity and relative weighted consistency) are used. Finally, we conduct extensive experiments on four microarray datasets and take other five widely used feature selectors as a comparison. The results indicate the effectiveness of the proposed ensemble selectors in obtaining stable and discriminant feature subsets.

TABLE III. THE NUMBER OF SELECTED FEATURES (AVG ± STD)

Dataset	FCBF		FCBF_H		FCBF_Q		FCBF_U		FCBF_F	
	#avg	*#std*	*#avg*	*#std*	*#avg*	*#std*	*#avg*	*#std*	*#avg*	*#std*
1	7.6	1.4	3.4	1.3	7.6	2.1	30.1	4.0	7.1	1.2
2	44.3	3.4	16.1	1.8	38.3	3.6	130.4	5.6	39.2	3.8
3	92.6	5.3	38.3	6.1	83.2	4.7	257.5	14.6	79.3	4.6
4	119.1	5.6	83.4	4.1	126.2	6.5	242.0	9.5	113.0	5.4

REFERENCES

[1] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816-824, 2002.

[2] A. Wang, N. An, G. Chen, L. Liu, and G. Alterovitz, "Subtype dependent biomarker identification and tumor classification from gene expression profiles," *Knowledge-Based Systems*, vol. 146, pp. 104-117, 2018.

[3] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, pp. 392-398, 2010.

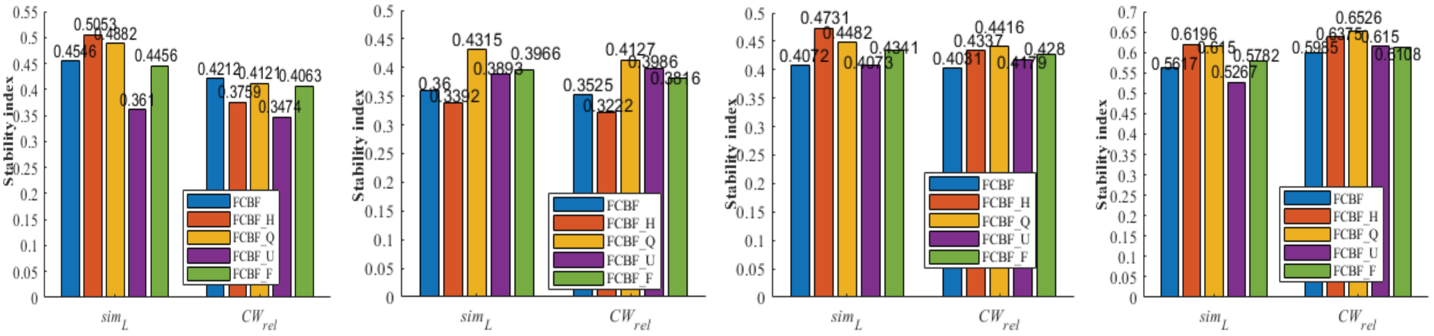
[4] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Information Fusion*, vol. 52, pp. 1-12, 2019.

[5] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 856-863.

[6] A. Wang, N. An, J. Yang, G. Chen, L. Li, and G. Alterovitz, "Wrapper-based gene selection with Markov blanket," *Computers in Biology and Medicine*, vol. 81, pp. 11-23, 2017.

[7] G. Brown, A. Pocock, M. J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27-66, 2012.

[8] P. Somol and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1921-1939, 2010.



(a) (b) (c) (d)

Fig. 2. Comparisons of stablity index. (a) *COLON*; (b) *DLBCL*; (c) *LEUKEMIA*; (d) *SRBCT*.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on January 17,2021 at 08:37:26 UTC from IEEE Xplore. Restrictions apply.