DOI 10.3233/BME-141206
IOS Press

Interacting gene selection via cooperative game analysis for cancer diagnosis

Xin Sun[a], Junyu Dong[a,*], Mantao Xu[b,c], Shengke Wang[a] and Cui Xie[a]
[a]*College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China*
[b]*School of Computing, University of Eastern Finland, Joensuu 111 FIN-80101, Finland*
[c]*School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200031, China*

Abstract. Microarray technologies offer practical diagnostic tools for cancer detection. One great challenge is to identify salient genes from the high dimensionality of microarray data that can directly contribute to the symptom of cancer. Interactions among genes have been recognized to be fundamentally important for understanding biological function. This paper proposes an interacting gene selection method for cancer classification by identifying useful interacting genes. The method firstly evaluates the interactivity degree of each gene according to the intricate interrelation among genes by cooperative game analysis. Then genes are selected in a forward way by considering both interactivity and relevance characters. Experimental comparisons are carried out on four publicly available microarray data sets with three outstanding gene selection methods. Moreover a gene set enrichment analysis is also performed on the selected gene subset. The results show that the proposed method achieves better classification performance and enrichment score than other gene selection methods.

Keywords: Medical diagnosis, cancer classification, gene selection, information theory, cooperative game analysis

1. Introduction

Gene expression analysis offers a practical and widely accepted diagnostic tool to facilitate the pathological diagnosis and cancer detection [1]. To help biologists and doctors obtain an accurate diagnosis, various data mining algorithms have been applied to analyze the gene expression microarray data. One of the greatest challenges is to identify salient small gene subset from thousands of genes in expression data that can directly contribute to the symptom of cancer. However, almost all the machine learning algorithms suffer from the inevitable “curse of dimensionality” caused by the high dimensionality of gene expression data and their small sample sizes. Investigations have shown that only a few genes are sufficient for cancer classification [2]. Therefore feature selection algorithms in the fields of machine learning are introduced to select the most informative gene subset for cancer diagnosis. It brings lots of benefits to bioinformatics, such as acquiring a better understanding of interaction among genes, reducing the diagnosis cost, and facilitating data visualization.

[*] Corresponding author: Junyu Dong, College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China. Tel.: 0532-66781729; Fax: 0532-66782300; E-mail: dongjunyu@ouc.edu.cn.

The most important question is to find the most relevant gene subset. To date, different machine learning approaches have been employed in gene selection [3]. Typically these methods fall into three types: embedded, wrapper and filter methods. The drawback of the embedded and wrapper methods is the high computational complexity and less generalization of the selected subset. In practice, filter selection methods have much lower computational complexity because it is independent from the learning algorithms. Specifically, information theory based gene selection methods achieves excellent performance and has drawn more and more attention (e.g. [3–6]). And it has been widely used to construct gene networks recently [7].

Researchers have found that genes are grouped according to similar cellular function [8], indicating that genes function normally in some small groups [8,9] where genes are highly interactive and each gene cannot function well apart from others. For example, recent research work (Wang et al. [10]) argues that a good informative gene should have a large number of directly connected genes in the co-expression network. However, most current selection methods based on information theory discard the genes that are highly correlated to the chosen ones, although they are relevant to the disease [5,11]. For instance, information theoretic methods usually introduce a 'Redundancy-Eliminate' criterion to delete the redundant genes [12,13]. It is likely to destroy the gene subset that have strong discriminatory power together but are weak as individuals [14]. This disadvantage is fatal to disease diagnosis and recognition [5]. Our previous research [5] indicates that many useful intrinsic gene groups (or groups of interactive genes) in high-dimensional gene expression data are helpful to improve recognition performance in cancer diagnosis. This work will tackle the above problem by introducing a cooperative game analysis method to evaluate the interactivity level of every gene according to the intricate and intrinsic interrelation among genes. Then a forward gene selection algorithm is proposed both considering the interactivity and relevance characters of every gene.

2. Entropy-based gene correlation analysis

In the following, some information-theoretic measurements [5] will be introduced for gene relevance, interaction and redundancy analysis.

2.1. Relevance

The more relevant gene means that the gene contains more information about the disease. The relevance of gene g to the disease can be calculated by mutual information.

$$R(g) = I(g; disease) \tag{1}$$

Because mutual information measurement tends to favor genes with more values, the continuous expression values of gene should be discretized into an equal number of values. As suggested by Ding and Peng in [15], each gene will be normalized firstly so that its mean and standard deviation are zero and one, respectively; then the continuous attributes will be discretized into three partitions, i.e., (−∞, −0.5], (−0.5, 0.5) and [0.5, +∞).

2.2. Interaction

Interaction implies each gene in the interaction group cannot work or survive apart from others. One typical example is the XOR problem, i.e., two variables which are useless alone can be useful together [13]. Suppose two interacting genes g_i and g_j, the relevance between g_i and disease can be increased conditioned by g_j. Then the following definition is given: two genes g_i and g_j will fall into an interaction relationship if the following form is satisfied.

$$I(g_i, disease \big| g_j) > I(g_i; disease) \tag{2}$$

2.3. Redundancy

Besides relevant and interacting genes, most of the genes in the microarray data can be treated as redundancy or irrelevance. A gene will be redundant if its relevance to the disease can be reduced by the appearance of any other genes. Formally, gene g_i will redundant with gene g_j if

$$I(g_i, disease \big| g_j) \le I(g_i; disease)\,. \tag{3}$$

This work focuses on retaining the most informative gene subset for cancer diagnosis. In the following sections, a cooperative game analysis method will be introduced to evaluate the interactivity degree of each gene and to output an informative gene subset.

3. The interacting gene evaluation and selection algorithm

3.1. Shapley value – a cooperative game theoretic method

Any subset of players $N=\{1, 2, ..., n\}$, including N itself, can form a coalition [16]. A coalitional game is a pair (N, v) in which N is a finite set of players, indexed by i; and $v{:}2^N \rightarrow R$ associates with each coalition $K \subseteq N$, a real-valued payoff $v(K)$ that the coalition's members can distribute among themselves, satisfying $v(\emptyset)=0$.

The Shapley value yields a unique outcome in coalitional games to compute the powers of players [17]. The Shapley value is denoted as $\phi(v)$, where $\phi(v) \in R^n$ and $\phi_i(v)$ is the payoff to the ith player as shown in Eq. (4).

$$\phi_i(v) = \sum_{\pi \subset N} \Delta_i(K) \frac{|K|!(n-|K|-1)!}{n!} \quad \text{and} \quad \Delta_i(K) = v(K \cup \{i\}) - v(K) \tag{4}$$

where n is the number of players and the sum extends over all subsets π of N except player i.

3.2. Gene evaluation and selection algorithm

The Shapley value calculates the distribution of the importance of players in the voting game, which can be transformed into the context of gene evaluation to estimate the interactivity level of genes [18]. Every coalition can be regarded as a candidate gene subset of the final selected optimal gene set. The Shapley value provides a practical way to estimate the gene interactivity level corresponding to the contribution that it makes to each of the subset it belongs to.

For convenience, an interacting index $\psi(i, j)$ is defined as Eq. (5) to indicate the interaction relationship between genes g_i and g_j, e.g., $\psi(i, j)$=1 denotes that they interact with each other.

$$\psi(i,j)=\begin{cases}1, & I(g_j;disease\,|\,g_i)>I(g_j;disease)\\ 0, & else\end{cases} \tag{5}$$

Then the function $\Delta_i(K)$ can be redefined with gene evaluation by

$$\Delta_i(K)=\begin{cases}1, & \sum_{f_j\in K}\psi(i,j)\geq\frac{|K|}{2}\\ 0, & else\end{cases} \tag{6}$$

which indicates a gene will be crucial to winning the coalition K only if it interacts with at least half of the members. According to Ockham's razor principle, the reason of assigning the threshold value as 1/2 is that the majority can control the coalition in game theory. Because every gene coalition is possible to be a subset of the final selected gene set, it is an effective method to measure the impact of gene g_i by computing the proportion of winning coalitions under conditions of g_i to all possible coalitions [13].

In fact, the calculation of the Shapley value for each gene requires summing all possible subsets of genes as shown in the Eq. (4), which is impractical in gene evaluation. However, the number of genes falling into one interaction group is much smaller than the total number of genes in the microarray data [5]. So it is unnecessary to review all coalitions for genes, especially the large gene coalitions. Thus, a limit value ω is suggested as a bound on the coalition size. The Eq. (4) can be redefined as Eq. (7), where Π_ω is the set of subsets of gene set $G\backslash i$ limited by ω.

$$\phi_i(v)=\sum_{\pi\subset\Pi_\omega}\Delta_i(K)\frac{|K|!(n-|K|-1)!}{n!} \tag{7}$$

In order to select an optimal gene subset in which genes are relevant to the disease and interact with each other, an interacting gene evaluation and selection procedure is designed to pick out the most informative genes. The pseudo codes are outlined as Algorithm 1.The process of the gene evaluation and selection algorithm is mainly composed of two stages: evaluation and selection.

Stage 1: An evaluation process based on the Shapley value is firstly presented to calculate the interactivity of each gene as shown in Step 2-5. The Shapley value can be regarded as a metric estimating the interactivity of every gene based on the complex structures among themselves.

Stage 2: Then a gene selection process based on mutual information is introduced to handle the gene selection problem. The Shapley values calculated in the first stage are used for regulating the relative importance of every gene for cancer diagnosis.

Algorithm 1

Interacting Gene Evaluation and Selection Algorithm (IGES)

```
Input: A microarray data set $O$ with gene set $G$ and cancer types $C$; User-specified threshold $k$.
Output: A set of selected genes $GS$.
Initialize gene set $GS=\emptyset$;
For each gene $g\in G$ do
    Create all coalitions set $\{\pi \mid |\pi|<\omega\}$ over $G\backslash g$;
    Calculate the Shapley value $\phi_g(v)$;
End
Compute the relevance $R(g)$ for each gene $g\in G$;
Compute the criterion $J(g)= R(g)\times \phi_g(v)$ for any $g\in G$;
Let $GS$ contain the first $k$ genes with the largest $J(g)$;
Return $GS$.
```

4. Experiments and results

In this section, comprehensive experiments are performed to compare the IGES method with three typical gene selection algorithms: mRMR [12], ReliefF [19], and SAM [20]. The prediction accuracy is used as metric by employing the Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) classifiers. Genes are always grouped together in an interaction group because they work together as a unit to produce proteins and achieve a particular biological function. Thus this work also analyzes the gene subset selected by the four methods in terms of gene set enrichment, so as to see whether the proposed method could find more biological meaningful genes.

4.1. Dataset descriptions

Experiments are carried out on four publicly available gene microarray data sets which are often used in literatures [5]: (1) Breast cancer dataset with 24481 genes of 97 patients, 46 of which have developed distance metastases within 5 years and the rest 51 remained healthy after their initial diagnosis. (2) Lung cancer dataset with 12533 genes contains 181 tissue samples (31 malignant pleural mesotheliomas and 150 adenocarcinomas). (3) Prostate cancer dataset has 52 prostate tumor samples and 50 non-tumor prostate samples with around 12600 genes. (4) Childhood ALL dataset consists of 60 childhood acute lymphoblastic leukemia samples, that is, 13 mercaptopurine alone, 21 high-dose methotrexate, 16 low-dose methotrexate and 10 high-dose methotrexate, with 8280 genes.

4.2. Prediction performance

For the sake of convenience in comparison, the experiment ranks all the genes and generates gene subsets by picking the top k genes, where k=1… 50. And the best subset will be chosen when it gets the highest accuracy for the classifier. Table 1 lists the statistics of the top classification accuracies with SVM and KNN classifiers. The number of selected genes is also recorded following the accuracy in Table 1. It can be seen that IGES performs better than others with both classifiers in most cases. In order to avoid the impact of the scarcity of data, the average accuracies for every selector are also shown in the "Ave." row. For example, the average accuracies of IGES with the two classifiers are 94.09% and 93.80% respectively, which are much better than others.

Table 1

The comparison of classification accuracies with SVM and K-NN

	SVM								**K-NN**							
No.	IGES		mRMR		ReliefF		SAM		IGES		mRMR		ReliefF		SAM	
	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#	Acc.	#
1	**94.73**	38	89.67	36	80.67	38	82.22	26	**92.44**	16	**100.0**	37	73.56	33	82.56	16
2	**100.0**	7	99.44	8	99.44	15	**100.0**	16	**100.0**	11	97.00	10	**100.0**	33	**100.0**	16
3	**98.63**	11	97.09	6	95.09	21	94.09	23	**99.50**	12	81.67	34	95.27	24	94.00	17
4	**83.00**	11	81.67	33	71.67	18	70.00	30	83.27	22	**91.59**	22	66.67	32	70.00	28
Ave.	**94.09**	17	92.70	23	84.41	23	86.14	25	**93.80**	15	85.78	12	81.75	27	83.59	17

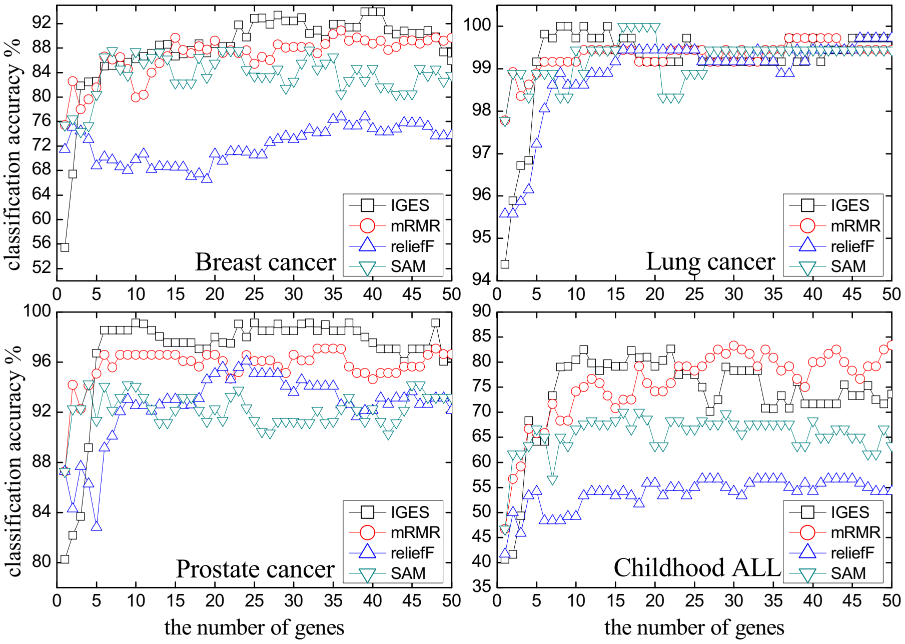

Fig. 1. Average accuracies vs. numbers of selected genes on the four cancer datasets.

To further verify the efficiency of the proposed method, the classification accuracies against the number of genes are compared as shown in Figure 1. The mean value of accuracies with the two classifiers is calculated in order to reduce the bias of a gene assessment based on a specific classifier. It can be seen from Figure 1 that the accuracies of IGES are very low with the first few selected genes. The reason is that IGES does not choose the first few genes with the maximal relevance to the disease. Nevertheless IGES becomes more excellent than the others after selecting a certain number of interacting genes.

4.3. Gene set enrichment analysis

As suggested by literature [5], gene set enrichment analysis is performed on the selected gene subset using GSEA software [21]. Cancer related gene sets are identified for four cancer types (256 of Breast

Table 2

The number of gene sets with FDR<0.25 or P-value < 0.001 in the GSEA

Data set	FDR < 0.25				P-value < 0.001			
	IGES	mRMR	ReliefF	SAM	IGES	mRMR	ReliefF	SAM
Breast Cancer	14	16	6	3	26	21	6	8
Lung Cancer	8	3	1	3	7	7	4	6
Prostate Cancer	3	2	3	3	6	2	4	3
Childhood ALL	2	1	0	1	5	4	1	3

Cancer, 186 of Lung Cancer, 67 of Prostate Cancer and 177 of Childhood ALL) from MSigDB [21] respectively. And the first thirty genes selected by the four selection algorithms are supplied to the GSEA software. Table 2 shows the number of gene sets that are related to cancer with false discovery rate (FDR) < 0.25 estimated by the GSEA. GSEA can also estimate the statistical significance (nominal P-value) of the Enrichment Score by using an empirical phenotype-based permutation test procedure, which preserves the complex correlation structure of the gene expression data. A gene set will be considered significantly enriched if the statistical significance (p-value) of its enrichment score is below the threshold. Table 2 also records the number of gene sets that show statistically significant relevance to cancer diagnosis at the selected genes. It can be seen from the results that the genes selected by the proposed IGES method is more enriched in gene sets that are related to cancer.

5. Conclusion

Genes are not acting in isolation but rather in gene groups for biological functions. This work focuses on identifying informative gene subset from the high dimensionality of microarray data that can directly contribute to the symptom of cancer. A novel gene evaluation and selection method is presented for cancer diagnosis by retaining interacting genes. Its primary characteristic is that the interactivity degree of each gene is firstly calculated by cooperative game analysis. So the optimal gene subset is selected by considering both interactivity and relevance of every gene. It can be found that excellent prediction performances are achieved by selecting the essential genes using IGES. Moreover, a gene set enrichment analysis on the selected gene subset is performed. The results show that the gene subset selected by IGES method achieves higher enrichment score than other gene selection methods.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61271405, 61301241), China Postdoctoral Science Foundation (No. 2014M551962), and Fundamental Research Funds for the Central Universities (No. 201413020).

References

[1] C.S. Cooper, Applications of microarray technology in breast cancer research, Breast Cancer Research **3** (2001), 158.

[2] W. Li and Y. Yang, How many genes are needed for a discriminant microarray data analysis, Methods of Microarray Data Analysis, 2002, 137–150.

[3] J. Xu, L. Sun, Y. Gao and T. Xu, An ensemble feature selection technique for cancer recognition, Bio-medical Materials and Engineering **24** (2014), 1001–1008.

[4] L. Sun and J. Xu, Feature selection using mutual information based uncertainty measures for tumor classification, Bimedical Materials and Engineering **24** (2014), 763–770.

[5] X. Sun, Y. Liu, D. Wei, M. Xu, H. Chen and J. Han, Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis, Journal of Biomedical Informatics **46** (2013), 252–258.

[6] X. Sun, Y. Liu, M. Xu, H. Chen, J. Han and K. Wang, Feature selection using dynamic weights for classification, Knowledge-Based systems **37** (2013), 541–549.

[7] A. Reverter and E.K.F. Chan, Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks, Bioinformatics **24** (2008), 2491–2497.

[8] D. Ruano, G.R. Abecasis, B. Glaser, E.S. Lips, L.N. Cornelisse, A.P.H. de Jong et al., Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability, American Journal of Human Genetics **86** (2010), 113–125.

[9] J.P. Demuth, T. De Bie, J.E. Stajich, N. Cristianini and M.W. Hahn, The evolution of mammalian gene families, PloS One **1** (2006), e85.

[10] Z. Wang, F.A. Lucas, P. Qiu and Y. Liu, Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection, BMC Bioinformatics **15** (2014), 153.

[11] C.-H Zheng, Y.-W. Chong and H.-Q. Wang, Gene selection using independent variable group analysis for tumor classification, Neural Computing & Applications **20** (2011), 161–170.

[12] H. Peng, F. Long and C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), 1226–1238.

[13] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen and X. Liu, Feature evaluation and selection with cooperative game theory, Pattern Recognition **45** (2012), 2992–3002.

[14] I. Guyon and E. Andr, An introduction to variable and feature selection, Journal of Machine Learning Research **3** (2003), 1157–1182.

[15] C. Ding and H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology **3** (2005), 185–205.

[16] J.W. Friedman, Game theory with applications to economics, Oxford University Press, New York, 1990.

[17] Q. Guo and M. Zhang, Implement web learning environment based on data mining, Knowledge-based Systems **22** (2009), 439–442.

[18] X. Sun, Y. Liu, J. Li, J. Zhu, X. Liu and H. Chen, Using cooperative game theory to optimize the feature selection problem, Neurocomputing **97** (2012), 86–93.

[19] K. Kira and L.A. Rendell, A practical approach to feature selection, Proceedings of the Ninth International Workshop on Machine Learning, 1992, 249–256.

[20] V.G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proceedings of the National Academy of Sciences **98** (2001), 5116.

[21] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), 15545.