A Novel Microarray Gene Selection Method based on Consistency

Yingjie Hu, Shaoning Pang and Ilkka Havukkala
KEDRI, Auckland University of Technology
Private Bag 92006, Auckland
1020, New Zealand
E-mail: krw5824@aut.ac.nz; spang@aut.ac.nz; ilkka.havukkala@aut.ac.nz

Abstract

Consistency modeling for gene selection is a new topic emerging from recent cancer bioinformatics research. The result of classification or clustering on a training set was often found very different from the same operations on a testing set. Here, we address this issue as a consistency problem. We propose a new concept of performance-based consistency and a new novel gene selection method, Genetic Algorithm Gene Selection method in terms of consistency (GAGSc).

The proposed consistency concept and GAGSc method were investigated on eight benchmark microarray and proteomic datasets. The experimental results show that the different microarray datasets have different consistency characteristics, and that better consistency can lead to an unbiased and reproducible outcome with good disease prediction accuracy. More importantly, GAGSc has demonstrated that gene selection, with the proposed consistency measurement, is able to enhance the reproducibility in microarray diagnosis experiments.

1. Introduction

The advent of microarray technology has made it possible to monitor the expression levels for thousands of genes simultaneously, which can help clinical decision making in complex disease diagnosis and prognosis, especially for cancer classification. A major challenge with microarray research is how to find informative genes that can be used for successfully discriminating variables in relation to different conditions, such as healthy and diseased. In a typical microarray dataset, the number of genes normally far exceeds that of samples. For example, there are 78 samples vs. 24,482 genes in the breast cancer dataset of [1]. Furthermore, those genes usually include many redundant genes that can confuse a classifying algorithm. Therefore, selecting informative genes from a dataset is a fundamental task for microarray research.

The reliability of microarray data analysis has been disputed in recent scientific literature, because many impressive results of microarray experiments could not be reproduced. In our previous experiments, the results obtained from the operations, such as classification and clustering on the training set were often found very different from the results of the same operations on the testing set. For example, the training set from CNS cancer data [2] can get a performance of above 90% true positive (TP) accuracy for tumour classification, whereas the testing data only gets 70% of TP accuracy. This occurs, because the typical methods for gene selection use only a single criterion of distance measurement between patients and non-patients, and disregard consistency between the subsets of data with the selected genes.

This issue is here discussed, which is referred to the consistency problem. Moreover, it is also noticed in our experiments that selecting a set of proper genes can significantly reduce the inconsistency of microarray data experiment. Obviously, it is more interesting to discover a set of genes that enable a consistently good classification performance over different subsets of patients in the complete microarray data.

The remainder of this paper is structured as follows. Section 2 services to briefly review the related work on consistency concept, following the description of our proposed consistency concept and GAGSc method. In section 3, the experiment setup and results are provided. The final section presents discussions and conclusions.

2. Consistency concept and GAGSc method

2.1. Related work

Probabilistic consistency analysis for gene selection method [3] is a recent novel approach that focuses on analyzing the common genes selected from two datasets. Consistency in this method is defined as follows:

Suppose two microarray datasets D_a and D_b targeting the same bioinformatics task, each having same number of genes. r is a ranking function generating two lists of sorted genes from the two datasets. Let s top-ranked genes in each case be selected and denoted by S_a and S_b. Then, the consistency C of this dataset is given by:

$$C(r, s, D_a, D_b) = |S_a \cap S_b| \qquad (1)$$

Consistency C is thus the number of genes in common between two datasets, and depends on ranking function, data and the number of selected genes [3].

Mukherjee and his colleagues have applied this probabilistic consistency to their data-adaptive (DA) gene selection method. In their gene selection process, the result of probabilistic consistency obtained from top-selected genes is used for optimizing the test statistics function. After hundreds of iterations, an optimized statistic function can be achieved based on consistency.

However, it is not clear to what extent the selected informative are related to the consistency of final classification performance, *i.e.* the performance of classification over individual sampling subsets of a complete dataset may still have a very inconsistent result, even though the common-gene method is employed.

2.2 Proposed consistency concept and GAGSc method

The idea of our approach is using the result of consistency obtained from an operation (*e.g.* classification, clustering, etc.) to find informative genes for a microarray dataset. For most microarray datasets, there tends to be no agreement on which genes are highly differentially expressed, and consequently it is difficult to measure the reliability of any gene selection method. In practice, the performance of an operation over microarray data is a straightforward criterion for measuring the outcomes of microarray experiments. Our new solution is based on optimizing computation which takes consistency into account.

The definition of our proposed consistency concept is as follows:

Consider a dataset pertaining to a bioinformatics task (two classes) and denoted by D. The dataset D consists of n samples with m genes, and all samples belong to two classes (e.g. class 1 or class 2). D_a and D_b are two subsets of D obtained by random subsampling, and serve as training and testing datasets, respectively.

$$D = D_a \bigcup D_b \quad \& \quad D_a \bigcap D_b = \varnothing \qquad (1)$$

Given a base function $\boldsymbol{F}$ over D, and a gene selection function f_s over D_a, the consistency of dataset D can be calculated as

$$C(\boldsymbol{F}, f_s, D) = |P_a - P_b| \qquad (2)$$

where P_a and P_b are the outcome of the function $\boldsymbol{F}$ on D_a and D_b,

$$P_i = \boldsymbol{F}(f_s(D), D_i) \mid i = \text{a, b}. \qquad (3)$$

Base function $\boldsymbol{F}$ can be any of various data processing models, such as clustering function, feature extraction function, classification function, *etc.*, it determines the feature space on which the consistency is based on. In the concept of consistency based on performance, $\boldsymbol{F}$ is set as one type of classification function.

As $\boldsymbol{F}$ is assigned as a classification function, the above fundamental consistency definition Eq. (2) can be extended as a definition of consistency in terms of classification performance,

$$C(\boldsymbol{F}, f_s, D) = |\boldsymbol{F}(f_s(D), D_a, D_b) - \boldsymbol{F}(f_s(D), D_b, D_a)| \qquad (4)$$

where $f_s(D)$ specifies D as the dataset for gene selection. D_a in the first term of Eq.(4) is assigned for classifier training, and D_b is for testing. The second term of Eq. (4) specifies a reversed training and testing position for D_a and D_b, respectively.

The key idea of Genetic Algorithm Gene Selection method in terms of Consistency (**GAGSc)** proposed in this work is using the result of consistency in terms of performance obtained from an operation (e.g. classification or clustering) to find a small set of informative genes for a microarray dataset. Meanwhile, an evolutionary function is employed for selecting candidate genes. Two genetic operators, mutation and crossover are applied to this evolutionary function for optimizing the gene selection function.

Given a dataset D, a list of genes S, and an operation function F_{sc} (e.g.classification), the optimized function performing GA method is expected to achieve:

$$f_s^* = \arg\min_{f_s \in \mathcal{F}} C(F_{sc}, S, D) \qquad (5)$$

where $\mathcal{F}$ refers to a family of evolutionary gene selection functions, F_{sc} and f_s refer to the function of computing consistency under the condition of gene selection and gene selection function, respectively.

The GAGSc algorithm can be simply summarized into the following steps:

Table 1. Summary of microarray and proteomics datasets used for experiments

Data name	Class 1 vs. Class 2	Number of Genes	Training Samples (class 1/2)	Validation Samples	Ref.
Lymphoma	Diffused large B cell lymphoma vs. other types	4026	(42/54) 96	-	[4]
Leukaemia	ALL vs. AML	7129	(27/11) 38	34	[5]
CNS tumour	Survivor vs. Failure	7129	(21/39) 60	-	[2]
Colon Cancer	Normal vs. Tumour	2000	(22/40) 62	-	[6]
Ovarian	Cancer vs. Normal	15154	(91/162) 253	-	[7]
Breast Cancer	Relapse vs. Non-Relapse	24482	(34/44) 78	19	[1]
Lung Cancer	MPM vs. ADCA	12533	(16/16) 32	149	[8]
Esophageal Cancer	Non-responder vs. Responder	859	(15/12)27	15	[9]

1. Split all genes of dataset D into ρ segments based on their mean value (note that ρ is a pre-specified number).
2. Randomly select one gene from each of ρ segments, respectively to form the initial candidate gene set that is denoted by S.
3. Apply the operation function F_{sc} to the dataset containing those genes listed in S, and compute the consistency C by Eq. (4).
4. Perform gene selection function f_s on S to get a new generation of genes S', and compute the consistency C'.
5. If $C' > C$, then $C = C'$ and $S = S'$.
6. Repeat Steps 3-5 for N generations. N is a given number (usually >= 200) for determining how many generations are used in this case.
7. Output the finally selected genes.

The optimized gene selection method is obtained after N generations based on the best consistency performance.

3. Experiments

3.1. Datasets

The proposed GAGSc method for selecting informative genes is applied to seven well-known benchmark cancer microarray datasets and one proteomics dataset. Table 1 summarizes the eight datasets used for gene selection in our experiment.

3.2. Unbiased verification policy

In most previous microarray data analysis work, sampling method is employed mainly for classification procedure, but not for gene selection procedure. Such methods produce the classification results eventually with bias, because the informative genes are selected from the whole dataset and not well estimated in terms of the generalization error. In practice, testing data is blind in real biology experiment, thus should not be allowed to be included in either gene selection or classification modelling. Therefore, the bias occurring in gene selection may finally result in an unreplicable disease diagnosis performance.

A totally unbiased verification policy for microarray analysis should guarantee that no generalization error occurs in either gene selection or classification procedures. To this end, efficient data sampling method should be used in the two procedures to maximally decrease the generalization error. In other words, the classification also needs to employ the verification methods to estimate the bias error. In our gene selection experiment, we utilized such a totally unbiased verification scheme in which the classification accuracy is obtained from an independent testing subset.

3.3. Results

For clarity, the classification accuracies obtained by GAGSc method is summarized in Table 2, and the reported accuracies in the papers listed in Table 1 is added as well. Our proposed GAGSc method outperforms the published methods on four datasets, and the classification result on colon data is very close to the reported accuracy. However, the classification accuracies of three datasets (CNS, Breast and Esophageal) are significantly lower than the published. As discussed in introduction, many published classification results are not based on efficient validation schemes, which results in the experiments are unreplicable and too optimistic. However, the experimental results obtained by the proposed GAGSc method can be easily reproduced, because the totally unbiased validation scheme is applied in this study. These results suggest that reproducible prognosis is possible for only 4 or 5 of the 8 used benchmark datasets in Table 2.

Table 2. Classification accuracy comparison: GAGSc results vs. known results from literature (See Refs in Table 1.)

Data	Classification accuracy	
	GAGSc	Publication
Lymphoma	**95.84%**	72.5%
Leukaemia	**94.12%**	85%
CNS Tumour	65.00%	83%
Colon Cancer	83.81%	87%
Ovarian	**98.80%**	97%
Breast Cancer	63.16%	94%
Lung Cancer	**91.28%**	90%
Esophageal Cancer	46.67%	93.3%

In our experiments, all data for validation is independent and never touched in the training process. Therefore, the selected informative genes are entirely fair to any given data for validation. Such a mechanism of gene selection might result in the bad performance in certain microarray datasets, which is due to their characteristics. However, the reported good results in published papers of these datasets are suspect.

The proposed GAGSc method has demonstrated that the consistency concept can be used for gene selection to solve the reproducibility problem in microarray data analysis. The main contribution of proposed GAGSc gene selection method is that it ensures the reliability and reproducibility of microarray data analysis experiments, and improves the disease classification performance as well.

According to our experimental results, we found that the classification performance could not be improved significantly after 200 generations. In general, using 20~30 (ρ) initially selected genes could produce the best and repeatable results.

4. Discussion and conclusions

We have briefly described a new gene selection method (GAGSc) in the proposed performance-based consistency theory. The main contributions of this study are: (1) The proposed consistency concept can be easily incorporated into more sophisticated gene selection systems to enhance the overall performance of microarray data analysis; (2) Using the proposed GAGSc method, the final selected informative genes can construct a better classifier for disease diagnosis in terms of prediction accuracies. The unbiased prediction accuracies on eight benchmark datasets obtained by GAGSc method in this study are very competitive to the reported results in literature. Note that some published prediction results are not validated on independent datasets, and thus remain suspect.

The findings of this study indicate that the proposed consistency concept in gene selection is a useful innovation in several areas. To further improve the gene selection methods based on the proposed consistency concept, it would be interesting to incorporate clustering in the pre-process stage of gene selection.

The huge computational complexity is one of the main limitations of the proposed GAGSc method. To alleviate computational complexity, cluster algorithms are intended to be used before GA search to find a certain number of clusters. Then these clusters can be used for determining the initial number of genes to be selected.

5. References.

[1] van't Veer, L., H. Dai, v.d.V. MJ, et al., "Gene expression profiling predicts clinical outcome of breast cancer". *Nature*. 415(6871): pp. 530-6, 2002.

[2] Pomeroy, S., P. Tamayo, M. Gaasenbeek, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression". *Nature*. 415(6870): pp. 436-442, 2002.

[3] Mukherjee, S. and S.J. Roberts. *Probabilistic Consistency Analysis for Gene Selection*. in *CSB*. IEEE Computer Society.pp. 487-488, 2004.

[4] Alizadeh, A.A., M. Eisen, R. Davis, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". *Nature*. 403(6769): pp. 503-11, 2000.

[5] Golub, T.R., D.K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science*. 286: pp. 531-537, 1999.

[6] Alon, U., N. Barkai, D.A. Notterman, et al. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. in *Proc Natl Acad Sci*.pp. 6745-50, 1999.

[7] Petricoin, E.F., A.M. Ardekani, P.J.L. Ben A Hitt, et al., "Use of proteomic patterns in serum to identify ovarian cancer". *Lancet*. 359: pp. 572-77, 2002.

[8] Gordon, G.J., R. Jensen, L.-L. Hsiao, et al., "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gege Expression Ratios in Lung Cancer And Mesothelioma". *Cancer Research*. 62: pp. 4963-67, 2002.

[9] Hayashida, Y., K. Honda, Y. Osaka, et al., "Possible Prediction of Chemoradiosensitivity of Esophageal Cancer by Serum Protein Profiling". *Clin Cancer Res*. 11(22): pp. 8042-47, 2005.