Gene Selection in Terms of Performance Based Consistency

Yingjie Hu, Shaoning Pang and Ilkka Havukkala
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 our recent cancer bioinformatics research. We found that the result of classification or clustering on a training set was often quite different compared to using a testing set. Here, we address this as a consistency problem. In practice, the inconsistency of microarray datasets prevents many typical gene selection methods working properly for cancer diagnosis and prognosis. In an attempt to deal with this problem, we propose a new gene selection method in terms of performance based on consistency.**

I. INTRODUCTION

The advent of microarray technology has made it possible to provide clinical decision support in complex disease diagnosis and prognosis area, especially in cancer treatment. In practice, DNA microarrays have been used for a variety of applications, such as separating tissue samples into two categories (e.g. healthy and diseased), predicting the clinical outcomes in response to cancer treatment and investigating the diagnosis outcomes in terms of the disease characteristics. It has been reported that the results of microarray experiments can be nearly 100% accurate [1, 2] Microarray technology is thus considered a revolution for studying all human disease, and effective therapy schemes for all complex diseases were predicted to be developed in a few years [3].

A major challenge with microarray research is how to find differentially expressed genes which can be used for effective 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, the breast cancer training dataset contains only 78 samples, but more than 24,000 genes [4]. Furthermore, those genes usually include many redundant genes that can confuse a classifying algorithm. Therefore, selecting highly relevant genes from the dataset is a fundamental task for microarray researchers.

Recently, the reproducibility of microarray studies has been debated in leading scientific journals, because it is seen as a crucial point in microarray technology application in medicine. The results of microarray experiments are difficult to reproduce due to the differences in the structures of microarray datasets produced in different laboratories. Recent work has attempted to establish effective validation rules to evaluate the reliability of microarray studies [5, 6]. We propose the concept of consistency to deal with this issue.

II. CONSISTENCY

A. Related work

Probabilistic consistency analysis for gene selection method [7] 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 $\mathcal{D}_a$ and $\mathcal{D}_b$ targeting the same bioinformatics task, each having same number of samples and 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 SET_a and SET_b. Then, the consistency C of this dataset is given by:

$$\text{C}(r, s, \mathcal{D}_a, \mathcal{D}_b) = |\ \text{SET}_a \cap \text{SET}_b| \tag{1}$$

Consistency C is the number of genes in common between the two sets, and depends on ranking function, data and number of selected genes [7]. Mukherjee and his colleagues have applied this probabilistic consistency to their gene selection method for selecting truly differentially expressed genes under various conditions. 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 the consistency.

B. Proposed method

The concept of consistency proposed by Mukherjee and his colleagues focuses on the common genes selected from two sampled datasets. However, it is not clear to what extent the “highly differentially expressed genes” selected in terms of consistency are related to the performance of the classification and clustering of the microarray data. In other words, the performance of classification or clustering over a dataset may not be improved, though their gene selection method based on consistency is applied. Motivated by this issue, we propose a new consistency concept in terms of performance.

The idea of our approach is using the result of consistency obtained from an operation (*e.g*. classification, clustering, etc.) to find differentially expressed 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.

Our gene selection process is an optimization, in which a genetic algorithm is employed. Mutation operation is applied in the genetic algorithm for optimizing the gene selection function. As one of the operations of genetic algorithms, mutation is an adaptive heuristic search algorithm based on the evolutionary idea. The main strength of mutation operation is that it helps solution convergence over successive generations towards the global optimum. Meanwhile, it also can provide a fast, effective and robust search, because new combinations implicitly contain more important information than using just the top-ranked individuals.

Consider a microarray dataset D pertaining to a bioinformatics task. Two subsets are randomly partitioned from D and denoted by D_a and D_b:

$$D = D_a \bigcup D_b \ \& \ D_a \bigcap D_b = \phi \qquad (2)$$

Given an operation function $\boldsymbol{f}$ (e.g. classification, clustering and etc.), and a gene selection function $\boldsymbol{S}$ over D_a and D_b, consistency C can be defined as the difference between the performance of an operation over dataset D_a and D_b:

$$\mathrm{C}(\boldsymbol{f}, \boldsymbol{S}, D_a, D_b) = | \mathrm{P_a} - \mathrm{P_b} | \qquad (3)$$

where $\mathrm{P_a}$ and $\mathrm{P_b}$ are computed by:

$$\mathrm{P}_i \ = \boldsymbol{f}(\mathrm{Set}_i, D_i) \mid i = \{\mathrm{a, b}\} \qquad (4)$$

where $\mathrm{Set_a}$ and $\mathrm{Set_b}$ are the selected genes obtained by the gene selection function $\boldsymbol{S}$ over D_a and that of D_b respectively, such that:

$$\mathrm{Set}_i = \boldsymbol{S}\,(\mathrm{D}_i) \mid i = \{\mathrm{a, b}\} \qquad (5)$$

The proposed gene selection algorithm is an optimizing process using an evolutionary function based on the consistency performance,

$$\boldsymbol{S}_i = \boldsymbol{E}(\boldsymbol{S}_{i-1}, \mathrm{C}_{i-1})\ ) \qquad (6)$$

The gene selection procedure can be summarized into the following steps:

1. Set an initial dataset Set_0 as a set of top-ranked genes from D using a typical gene selection function, e.g. t-test or SNR algorithm.
2. Apply the operation function $\boldsymbol{f}$ to Set_0 to calculate the consistency C_0:

$$\mathrm{C}_0 \ = \ |\boldsymbol{f}(\mathrm{Set}_0, D_a) - \ \boldsymbol{f}(\mathrm{Set}_0, D_b)| \qquad (7)$$

3. Mutation of gene selection. Use an evolutionary function to optimize the gene selection function.
4. Compute the consistency C′ based on the evolutionary gene selection function:

$$\mathrm{C}' \ = \ |\boldsymbol{f}(\mathrm{Set}', D_a) - \ \boldsymbol{f}(\mathrm{Set}', D_b)| \qquad (8)$$

where Set′ is obtained via the optimized gene selection function.

5. If $\mathrm{C}' > \mathrm{C}_0$, the consistency is improved due to the evolutionary gene selection function. Then, we set C_0 = C′ and Set_0 = Set.
6. Repeat Steps 3-5 for N generations.
7. Output the best Set obtained.

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

III. EXPERIMENTS

We evaluated our proposed concept for gene selection on four well-known benchmark microarray datasets, and compared the results with three existing methods, including the t-test, SNR (Signal Noise Ratio) and Data-adaptive [8] methods. The four datasets are:

1. Golub's ALL-AML Leukaemia training data [9] containing 38 bone marrow samples (27 ALL and 11 AML). The dimensionality of this microarray data is 7,129 probes from 6817 human genes.
2. Colon Tumor dataset [10]. This dataset consists of 62 samples collected from colon-cancer patients; 40 samples are labeled as cancer and 22 are labeled as normal. Only 2,000 genes out of total 6,500 genes are selected into the dataset based on the confidence in the measured expression levels.
3. Lymp4026 data [11] contains the expression levels of 4026 genes across 47 samples in lymphoma patients.. Among them, 24 samples are from "germinal center B-like" group while 23 are "activated B-like" group.
4. Central Nervous System (CNS) cancer data [12]. This

dataset contains 60 patient samples in which 21 are survivors and 39 are non-survivors.

The number of top-ranked genes for function ***f*** in our experiment is set as 50. The number of selected top genes is usually in relation to the performance of the operation over datasets. Fewer genes may make the operation unreliable, whereas too many genes selected might introduce noise to the experiment. Previous studies indicate that a few dozen to a few hundred top-ranked genes can efficiently classify the different disease patterns in most microarray experiments. For example, the best classification result for the Lymphoma data occurs using approximately 50 to 200 top genes [13].

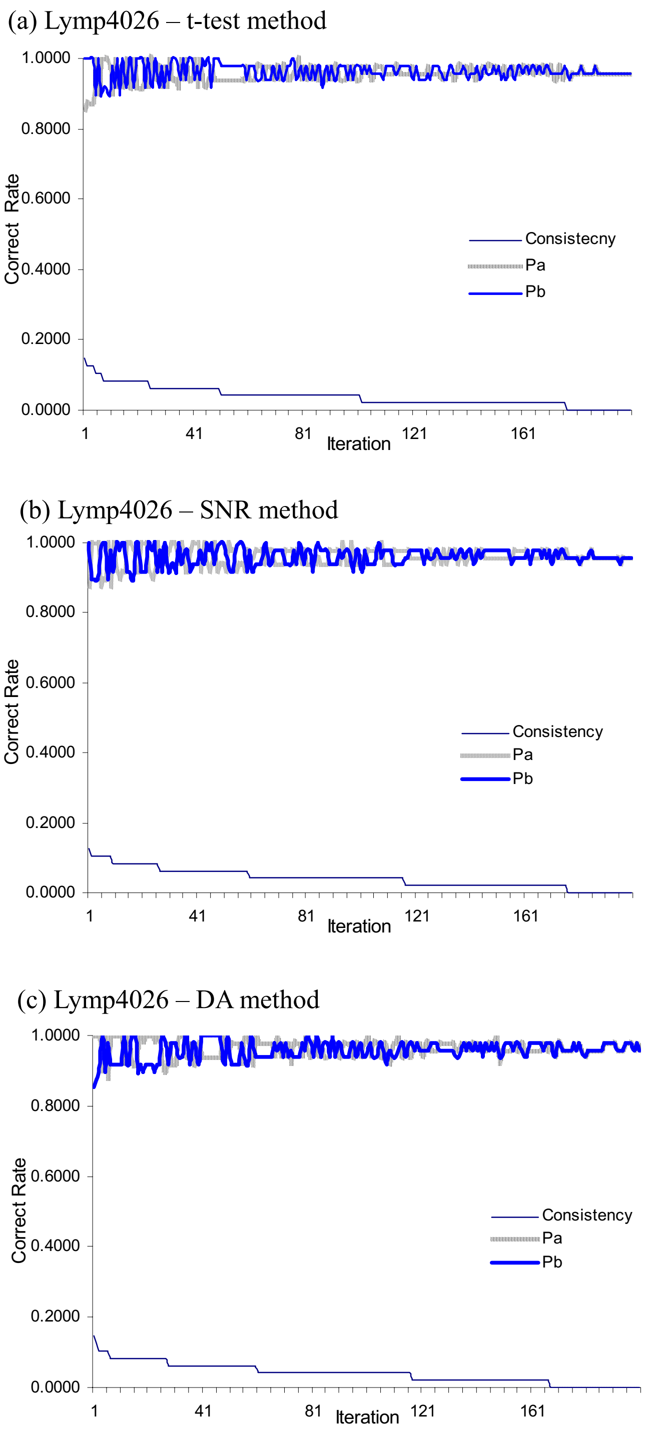
We have applied three gene selection methods to do the initial gene selection over the above four cancer microarrays, respectively. K-nearest neighbor (KNN) algorithm is used as the operation function ***f*** to do classification over four microarray datasets. The consistency is evaluated in terms of mean for 200 runs by Eq. (3), and its value ranges from 0 to 1. The best value of consistency should be 0, which means the accuracies of classification of the two subsets D_a and D_b are the same. If the value of consistency is 1, the classification accuracies of D_a and D_b are totally different. The value of performance P_a and P_b also ranges from 0 to 1, in which 1 means all the samples are 100% classified into the correct classes whilst 0 is the poorest accuracy.

Data Consistency Method	Leukaemia	Colon	Lymp4026	CNS
t-test	0.0529 ± 0.0400	0.1227 ± 0.0916	0.0411 ± 0.0314	0.0863 ± 0.0710
SNR	0.0455 ± 0.0383	0.1123 ± 0.0853	0.0421 ± 0.0299	0.0985 ± 0.0744
Data-adaptive	0.0674 ± 0.0581	0.1232 ± 0.0943	0.0379 ± 0.0249	0.1275 ± 0.0942

Table. 1. Comparisons of three methods in terms of performance based on consistency. The value of consistency is represented by mean ± standard deviation value.

Table 1 shows different datasets have very different inherent consistency values. The consistency of Lymp4026 and Leukaemia is significantly better than that of Colon and CNS data. In contrast, Colon data has the highest inconsistency that is nearly three times higher than the most consistent dataset (Lymp4026).

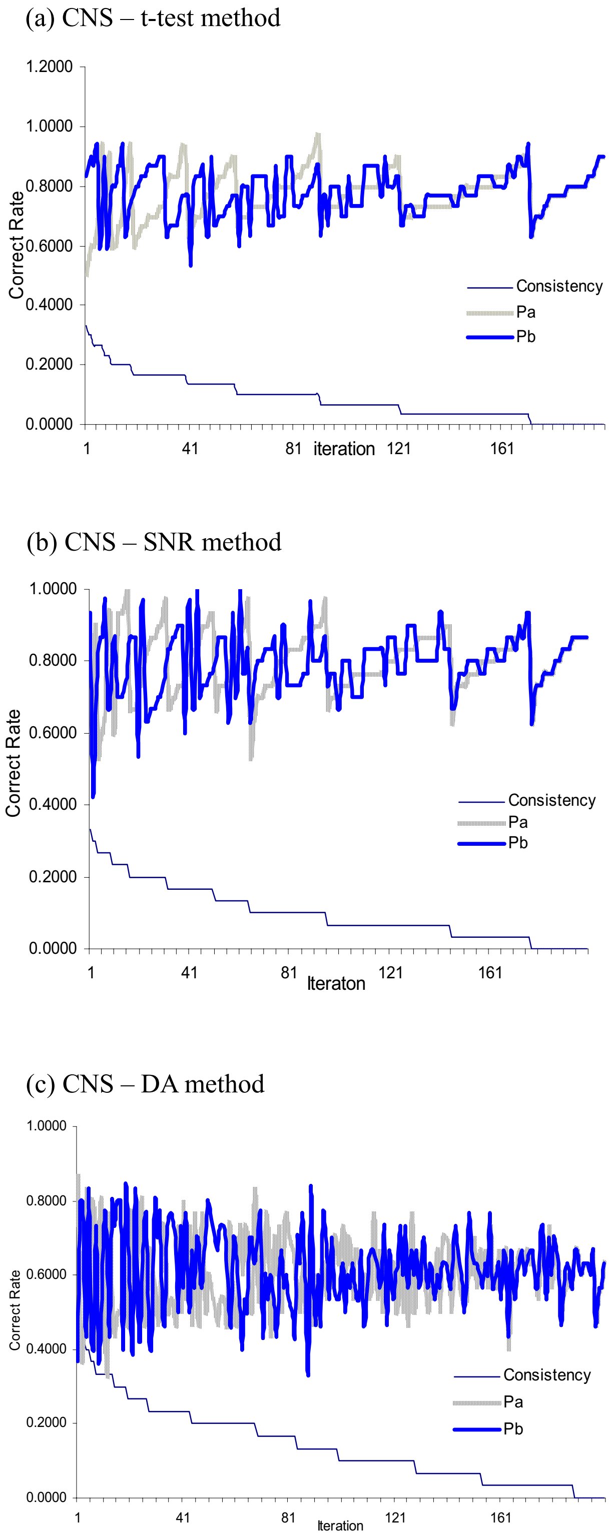
In addition, the consistency is seen also varying over different gene selection methods. SNR method outperforms the other methods on three of the four datasets, while Data-adaptive method wins in the fourth dataset. Among these datasets, the best consistency occurs when Data-adaptive method is used for gene selection on Lymp4026 dataset. Thus, the consistency performance also depends on the gene selection method even on the same dataset.


(a) Lymp4026 – t-test method

(b) Lymp4026 – SNR method

(c) Lymp4026 – DA method

Fig 1. Comparison of three methods (t-test, SNR, Data-adaptive (DA) gene selection methods) on Lymp4026 data in terms of performance based on consistency. X axis represents the iteration times, and Y axis is the value of consistency via Eq. (3), classification accuracy of D_a and D_b (P_a and P_b) respectively, which calculated by Eq. (4).



(a) CNS – t-test method

(b) CNS – SNR method

(c) CNS – DA method

Fig 2. Comparison of three methods on CNS dataset in terms of the performance based on consistency.

As seen in Fig 1, for all three gene selection methods, the consistency of Lymp4026 data in terms of performance is stable, with lower values (closer to zero). Correspondingly, it displays better average classification performance (P_a and P_b) of Lymp4026 data. In contrast, in Figure 2 (a)–(c), CNS data shows more variability in both consistency and performance. Interestingly, we can see that better consistency is related to the high performance. In Fig (2), more variability is shown in the value of consistency, P_a and P_b using DA method, while t-test and SNR methods show less variability.

The experimental results show the consistency and performance of microarray data can be improved by using appropriate methods for gene selection, though different data has different inherent consistency. We also find there is no gene selection method which is always the best for the above four microarray data in terms of consistency and performance.

IV. DISCUSSION AND CONCLUSIONS

We briefly outlined a new approach based on consistency in terms of performance in the assessment of existing gene selection methods for various microarray datasets. This approach is straightforward to show the strength of different gene selection algorithms, which can assist researchers to choose proper methods for doing various classifications and clustering over particular microarray data.

We described a new gene selection method in terms of performance based on consistency. A genetic algorithm for the given initial gene selection function leads to an optimized gene selection method from a family of evolutionary functions. The performance of classification and clustering over the dataset based on consistency is improved simultaneously with the set of selected informative genes.

In a set of experiments, we noticed that the results from different existing gene selection methods on the same data are quite similar. However, this is not sufficient to show which method fits best for the operation over a particular microarray data. Thus, we aim to revise our gene selection method in future work, which can be summarized into the following points:

1. The consistency C could be computed as:

$$C(\boldsymbol{f}, \boldsymbol{S}, D_a, D_b) = |\boldsymbol{f}(\mathrm{Set}_a, D_a) - \boldsymbol{f}(\mathrm{Set}_a, D_b)| \qquad (9)$$

Where Set_a is computed by:

$$\mathrm{Set}_a = \boldsymbol{S}\,(D_a) \qquad (10)$$

The consistency is independent on D_b, which ensures the testing data has no relation with the computed consistency. We are currently working on this extended method.

2. Only one operation over microarray datasets (KNN classification) was applied in our experiments. We will use other classification and clustering algorithms to do the validation of our proposed gene selection method.

3. Design a new scheme for sampling the raw microarray dataset. In general, it is fairly difficult to detect what is the best splitting ratio for a particular dataset sampling. We are looking into the use of other methods for simple but effective sampling of raw microarray data.

In conclusion, we have developed a new genetic algorithm approach for gene selection in terms of consistency. This method can be used in a variety of analysis tasks of high-dimensionality data, not only microarray data. Further development of the approach is in progress.

REFERENCES

[1]. 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, 2002. 359: p. 572-77.

[2]. Zhu, W., X. Wang, Y. Ma, et al. *Detection of cancer-specific markers amid massive mass spectral data*. in *Proc Natl Acad Sci U S A*. 2003.

[3]. Schena, M., *Microarray analysis*. 2002, New York: John Wiley & Sons.

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

[5]. Ransohoff, D.F., *Rules of evidence for cancer molecular marker discovery and validation.* Nature Reviews Cancer, 2004. 4: p. 309-314.

[6]. Ioannidis, J.P.A., *Microarrays and molecular research: noise discovery?* Lancet, 2005. 365: p. 453-455.

[7]. Mukherjee, S. and S.J. Roberts. *Probabilistic Consistency Analysis for Gene Selection*. in *CSB*. 2004. Stanford, CA, USA: IEEE Computer Society.

[8]. Mukherjee, S., S.J. Roberts, and M.J.v.d. Lann, *Data-adaptive test statistics for microarray data.* Bioinformatics, 2005. 00(00): p. 1-7.

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

[10]. Alon U, Barkai N, Notterman DA, 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*. 1999. USA.

[11]. Alizadeh AA, Eisen MB, Davis RE, et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. 403(6769): p. 503-11.

[12]. Pomeroy, S., P. Tamayo, M. Gaasenbeek, et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. 415(6870): p. 436-42.

[13]. Li, L., C.R. Weinberg, T. Darden, et al., *Gene Selection for sample calssification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. 17(12): p. 1131-1142.