Using Gene Clustering to Identify Discriminatory Genes with Higher Classification Accuracy

Zhipeng Cai*
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: zhipeng@cs.ualberta.ca

Lizhe Xu*
The Research Institute for Children-University of New Orleans
New Orleans, LA 70118, USA
Email: lizhe.xu@dhs.gov

Yi Shi, Mohammad R. Salavatipour, Randy Goebel and Guohui Lin†
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: ys3, mreza, goebel, ghlin@cs.ualberta.ca

***Abstract*— A single DNA microarray measures thousands to tens of thousands of gene expression levels, but experimental datasets normally consist of much fewer such arrays, typically in tens to hundreds, taken over a selection of tissue samples. The biological interpretation of these data relies on identifying subsets of induced or repressed genes that can be used to discriminate various categories of tissue, to provide experimental evidence for connections between a subset of genes and the tissue pathology. A variety of methods can be used to identify discriminatory gene subsets, which can be ranked by classification accuracy. But the high dimensionality of the gene expression space, coupled with relatively fewer tissue samples, creates the dimensionality problem: gene subsets that are too large to provide convincing evidence for any plausible causal connection between that gene subset and the tissue pathology. We propose a new gene selection method, clustered gene selection (CGS) which, when coupled with existing methods, can identify gene subsets that overcome the dimensionality problem and improve classification accuracy. Experiments on eight real datasets showed that CGS can identify many more cancer related genes and clearly improve classification accuracy, compared with three other non-CGS based gene selection methods.**

***Index Terms*— Microarray gene expression data; Gene clustering; Gene selection; Classification**

I. Introduction

DNA microarrays have provided the opportunity to measure the expression levels of thousands of genes simultaneously. Such an emerging technology enables the language of biology to be spoken in mathematical terms, yet abstracting useful information from a large volume of experimental microarray data remains challenging. One of the most common applications is to use microarrays for comparing the gene expression levels in tissues under different conditions, such as wild-type versus mutant, or healthy versus diseased. In general, among the thousands to tens of thousands of genes that are monitored simultaneously in multiple experiments, only a fraction of them are biologically relevant and can be identified as contributors to tissue samples' properties. These interesting genes are usually differentially regulated under experimental conditions, i.e, their expression levels are increased or attenuated, compared to the normal levels. Identifying these *discriminatory* genes is very important in many applications such as disease subtype discovery and profiling [1], [2], [3]. Other genes, such as house-keeping genes whose expression levels are largely unchanged under different conditions, are less important in providing information to downstream data analysis.

Microarray experiments are still expensive and an application typically requires a volume of data that may take several months or years to accumulate. Within one dataset, the number of samples is normally small, only in tens to hundreds, compared to the large number of genes monitored. Therefore, such a high dimensionality gene space has to be carefully managed in class prediction. Both dimensionality reduction and gene selection have been used for this purpose. This paper addresses the latter: to identify a subset of discriminatory genes that can be used for effective class prediction.

A variety of approaches have been proposed for gene selection. For example, Golub *et al.* [1] developed a measure of correlation that emphasizes the "signal-to-noise" ratio in using the gene as a predictor, and selected a number of top ranked genes as discriminatory genes. This ratio captures the basic rule of gene selection: that a discriminatory gene must have close expression levels in samples within a class, but significantly different expression levels in samples across different classes. Other approaches that adopt the same principle, with modifications and enhancements, include [4], [5], [3] and many others. Xiong *et al.* [6] select a subset of genes with a

*ZC and LX contributed equally to this work.
†To whom correspondences should be addressed.

Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)
0-7695-2727-2/06 $20.00 © 2006 IEEE

maximum classification accuracy through sequential (floating) forward selections (SFS, SFFS). Guyon *et al.* [7] suggest a gene selection method that uses support vector machines (SVMs) and recursive feature elimination (RFE).

It is of great interest that, on a common microarray dataset, different gene selection methods reported different subsets of genes, though they all achieved high classification accuracies [7], [5], [3]. One explanation has been that many genes have similar discrimination power, thus although biologically relevant, including them all would make some portion of the selected gene subset redundant for classification purpose. Consequently, gene selection methods that "order" genes differently will report different gene subsets. Nonetheless, identifying all correlated genes is equally important to classification, though a small subset of them might have reached 100% classification accuracy. On the other hand, when a subset of genes reaches 100% classification accuracy, it is often difficult to identify another discriminative gene by adding it to the subset.

In this paper, we address the issue of reporting all possibly correlated genes through combining gene clustering and gene selection in class prediction. Essentially, we allow only a limited number of genes from a cluster to be selected for classification purposes, where a cluster is a subset of genes that have very similar expression patterns across all samples. The output of such a clustered gene selection (CGS) method is a subset of genes that represent several clusters of similarly expressed genes. This way, by limiting the number of genes from a cluster, it leaves room for less correlated genes to be discovered, as well as some complementary genes that, individually, do not do well at separating the data.

The rest of the paper is organized as follows: After introducing some basics of gene selection, we present the details of the CGS method in Section II. Section III describes the experiments on combining CGS with two classifiers: k nearest neighbors (KNN) and a linear kernel support vector machine (SVM). Two most complex of the eight real microarray datasets that were used in the experiments are also described in the section. We discuss our results in Section IV. Specifically, we examine the impact of varying the number of clusters, the number of genes to be selected, and the distance measure in gene clustering (the Pearson correlation coefficient and the Euclidean distance). We also examine the quality of CGS, through comparisons to the non-CGS based gene selection methods Cho's [5], F-test [4], [2], and GS [3]; the differences in the performances of KNN- and SVM-classifiers when combined with CGS and non-CGS based Cho's, F-test, and GS are also examined. Section V summarizes our conclusions.

II. The Clustered Gene Selection Method

There are two challenges in microarray data classification. One is class discovery to define previously unrecognized classes; but this is not the focus of this paper. The other is to assign individual samples to already-defined classes. Methods such as the one in [7] can do the classification directly on the dataset by using all genes, though it might suffer from the dimensionality problem. Another approach is to select a subset of discriminatory genes and use only them as features for class prediction. In fact, one of the main purposes of gene selection is to identify biomarkers that can effectively predict the classes for samples. To this goal, a number of samples with known class labels are provided, which form the *training* dataset. The classifier built on the selected genes is tested on unlabeled samples and its performance is measured by the *classification accuracy*, which is defined as the number of correctly identified samples divided by the total number of testing samples. The leave-one-out (LOO) cross validation method is a process that uses one sample for testing and all the others for training, then the process is repeated for every sample in the dataset. Another popular cross validation is ℓ-fold, in which the whole dataset is partitioned into ℓ equal parts and, at one time, one part is used for testing and the other $\ell - 1$ parts for training. In our experiments, we used both cross validation methods, but chose to report only 5-fold average classification accuracy over 20 iterations of random partitions. The LOO cross validation results are included in Supplementary Materials.

Many existing gene selection methods are based on a gene scoring function that assigns a score for each gene, which approximates the relative discriminatory strength of the gene. Such gene scoring functions can be the classification accuracy of individual genes [6], or capture the basic rule that discriminatory genes are those being close at expression levels in intra-class samples but being significantly different in inter-class samples [4], [5], [3]. Among the latter category there are T-test and F-test based gene selection [4], [2], Cho's gene selection [5], and an improved version of Cho's called the GS method [3]. Here we adopt these three gene selection methods as our base methods. These methods generally return a number of top ranked genes, and their quality is measured by the classification accuracy of the classifier built on them. It is noticed that some correlated genes have very similar expression profiles. Consequently, they must have similar discrimination power in terms of classification, and once one is top ranked, the others will also be top ranked. However, using them all in building a classifier is redundant in that their discrimination power overlaps. Furthermore, we believe that when the number of selected genes is pre-specified, other genes with less discriminatory power would be excluded, and, as we later show, prevent improvements in classification.

Based on the above observations, we propose to do gene clustering before gene selection. Assume the training dataset consists of n genes and m samples. We apply a k-means clustering algorithm [8] to group the n genes into k clusters, for some k (to be discussed below). Essentially, k-means is a centroid-based clustering algorithm that partitions the genes into k clusters based on their pairwise distance, to ensure that intra-cluster similarity is high and inter-cluster similarity is low. We use both the Euclidean distance (`http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm`) and the Pearson correlation coefficient (`http://rana.lbl.gov/EisenSoftware.htm`) in our k-means experiments. At the same time, a gene scoring function is used to order the genes. With the cluster

Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)
0-7695-2727-2/06 $20.00 © 2006 IEEE

information for each gene, the CGS method scans the ordered gene list to pick up a total of L genes such that at most C genes from each cluster are selected. That is, when there are already C genes from a cluster, other genes belonging to the same cluster are simply skipped. Depending on the scoring function used, which can be Cho's, F-test, or GS, we refer to the resulting gene selection methods as CGS-Cho's, CGS-F-test, and CGS-GS, respectively. Subsequently, these L selected genes are fed to a KNN-classifier [3] and an SVM-classifier with a linear kernel (`http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox/`), to construct the class predictor. For example, CGS-GS-SVM refers to the SVM predictor built on genes selected by CGS-GS.

III. Experimental Results

We included three gene selection methods in our experiments: Cho's [5], F-test [2], and GS [3] (the GS1 in [3], another similar one GS2 is included in Supplementary Materials), from which we constructed three CGS based methods: CGS-Cho, CGS-F-test, and CGS-GS. We applied these six methods on eight real microarray datasets, among which the CAR dataset is probably the most complex, as it contains only 174 samples but in 11 classes, followed by the GLIOMA dataset as the second most complex. Here we report results of the KNN-classifier, with $k = 5$. (The Matlab source code for KNN-classifier is included in Supplementary Materials.) All experiments were conducted in the Matlab (`http://www.mathworks.com`) environment on a cluster of 32 2.33GHz CPUs. We focus on the report of our results mostly on the CAR dataset, with some results from the GLIOMA dataset. Only 5-fold cross validation results are reported. Complete results are not included due to space limit, but they are all available through Supplementary Materials.

A. Dataset Descriptions

The GLIOMA dataset contains 50 samples in four classes: *classic glioblastoma*, *nonclassic glioblastoma*, *classic anaplastic oligodendroglioma* and *nonclassic anaplastic oligodendroglioma*. This dataset is from Affymetrix U95av2 GeneChips. These classes have 14, 14, 7, 15 samples, respectively [9]. Each sample originally had 12,625 genes. We apply a standard filtering method to remove genes with minimal variations across the samples [9]. For this dataset, the intensity thresholds were set at 20 and 16000 units. That is, all hybridization intensity values less than 20, including negative hybridization intensity values, were raised to 20; and those higher than 16000 were shifted to 16000. Also, genes whose expression levels varied < 100 units between samples, or varied < 3 folds between any two samples, were excluded. After preprocessing, we obtained a dataset on 4,433 genes.

The CAR dataset contains in total 174 samples in eleven classes: *prostate*, *bladder/ureter*, *breast*, *colorectal*, *gastroesophagus*, *kidney*, *liver*, *ovary*, *pancreas*, *lung adenocarcinomas*, and *lung squamous cell carcinoma*, which have 26, 8, 26, 23, 12, 11, 7, 27, 6, 14, 14 samples, respectively [10]. This dataset is from Affymetrix U95a GeneChips. Each sample originally contained 12,533 genes. We preprocessed the dataset as described in [10] to include only those probe sets whose maximum hybridization intensity in at least one sample was 200; And all hybridization intensity values less than 20, including negative hybridization intensity values, were raised to 20, and the data were log transformed. After preprocessing, we obtained a dataset on 9,182 genes.

Descriptions of the other six datasets can be found in Supplementary Materials.

B. 5-Fold Cross Validation Classification Accuracies

In the discussion section, we report the experimental results on choosing the value for k in the k-means clustering algorithm and choosing the value for C, the maximum number of genes from each cluster. Those results support that $k = 100$ and $C = 1$ is one of the best settings that often achieves the highest classification accuracies on the GLIOMA dataset. For this reason, we chose to set $k = 100$ and $C = 1$ as defaults. Also, the default similarity measure in the k-means is the Pearson correlation coefficient (cf. Subsection IV.B). Figure 1 and 2 plot the 5-fold cross validation classification accuracies for the Cho's, F-test, GS, CGS-Cho's, CGS-F-test, and CGS-GS gene selection methods on the GLIOMA and CAR datasets, respectively, where results for the KNN-classifier and SVM-classifier are plotted separately. We also tested a gene selection method to randomly pick a subset of genes, and used its experimental results as a baseline. This method is denoted as *Random*. Subsequently, combined with a KNN-classifier and an SVM-classifier, we have Random-KNN and Random-SVM, respectively, and their performances are also plotted in the figures. We did not test the ℓ-fold cross validation for other values of ℓ in the current work.

The classification accuracies on all the eight datasets show that the CGS based gene selection methods outperformed their non-CGS based counterparts. The CAR dataset seems the most complex among the eight datasets, since it contains the largest number of genes and the most classes. Non-CGS based methods actually performed poorly as seen in the three plots in Figures 1 and 2. For example, in Figure 2, the average 5-fold classification accuracies are all below 50% when 20 genes or less are used. The CGS based methods performed much better and can reach above 80% when 20 genes are used. Further detailed quantitative improvements by the CGS method will be discussed in Section IV.C.

Figure 3 and Figure 4 show the standard deviations over 100 5-fold classification accuracies of all the seven gene selection methods, combined with KNN-classifier and SVM-classifier, on the GLIOMA and CAR dataset, respectively. The results show that the standard deviations of accuracies of all the methods do not have any significant difference within a dataset, though they are a little bit larger in the GLIOMA dataset than in the CAR dataset.

Regarding the baseline performance, the Random method was not bad and in fact in a few cases, the experimental results are even better than some of the non-CGS-based methods. This fact strongly suggests the dimensionality problem in the microarray datasets. Nonetheless, overall, both the non-CGS-based methods and Random performed much worse than the

Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)
0-7695-2727-2/06 $20.00 © 2006 IEEE

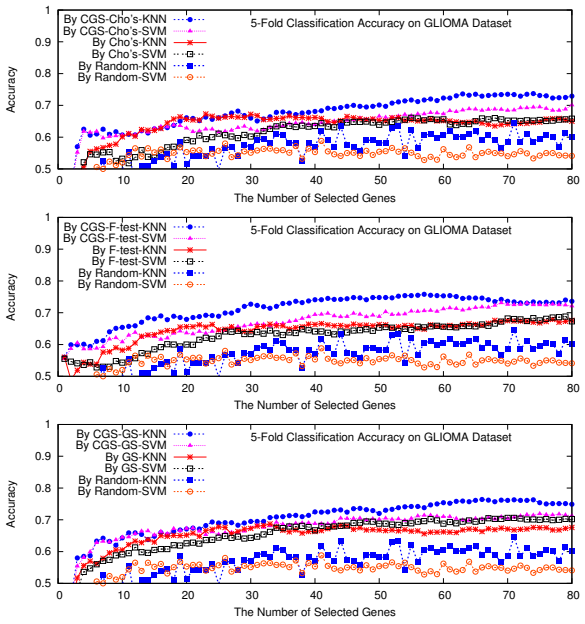

Fig. 1

5-FOLD CLASSIFICATION ACCURACIES OF ALL THE SIX GENE SELECTION METHODS, COMBINED WITH KNN-CLASSIFIER AND SVM-CLASSIFIER, ON THE GLIOMA DATASET.

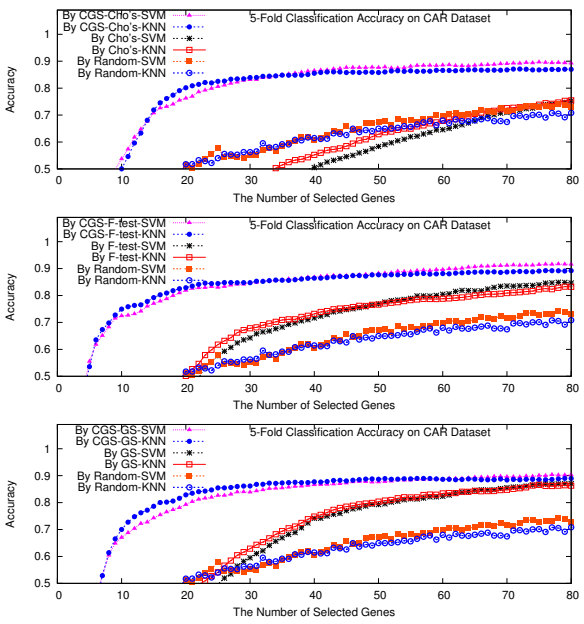

Fig. 2

5-FOLD CLASSIFICATION ACCURACIES OF ALL THE SIX GENE SELECTION METHODS, COMBINED WITH KNN-CLASSIFIER AND SVM-CLASSIFIER, ON THE CAR DATASET.

CGS-based methods, and these notably higher classification accuracies by the CGS-based methods demonstrate that the CGS-based methods are able to reduce the overfitting effect to identify more hidden discriminatory genes.

IV. DISCUSSION

A. Number of Clusters and Number of Genes Per Cluster

In the gene clustering stage, the k-means algorithm requires a manual input k that defines the number of clusters to be returned. On each dataset, the number of gene clusters affects the sizes of the resultant clusters, which in turn requires the tuning of C, the maximum number of genes selected from each cluster as discriminatory genes. Intuitively, a smaller k suggests a larger gene pool for a cluster, and therefore a larger value should be set for C. On the GLIOMA dataset, we have performed experiments using different combinations of k $(80, 90, \ldots, 140, 150)$ and C $(1, 2, 3, 4, 5)$. For each k, its quality is measured by the average 5-fold classification accuracy over five values of C, that is, a total $100 \times 5 \times 3 \times 2 = 3000$ classification accuracies, where 100, 3 and 2 indicate a hundred 5-fold training and testing, three CGS based methods and two different classifiers respectively (only the Pearson correlation coefficient was tested). Similarly, for each C, its quality is measured by the average 5-fold classification accuracy over eight values of k, that is, a total $100 \times 8 \times 3 \times 2 = 4800$ classification accuracies. All these average classification accuracies are plotted in Figure 5. These plots indicate that $C = 1$ is the best choice and $k = 90, 100, 110$ are equally good. This strongly suggests that, according to the expression

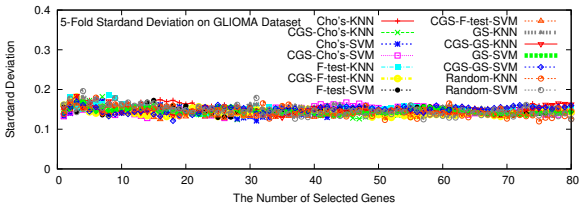

Fig. 3

STANDARD DEVIATION OVER 5-FOLD CLASSIFICATION ACCURACIES OF ALL THE SEVEN GENE SELECTIONS METHODS, COMBINED WITH KNN-CLASSIFIER AND SVM-CLASSIFIER, ON THE GLIOMA DATASET.

levels, genes can roughly be grouped into around 100 clusters, and that among the similarly expressed genes from a cluster, using one of them for building classifier is probably sufficient. We therefore set $C = 1$ and $k = 100$ as the default setting for the CGS based gene selection methods.

B. Distance Measure in k-Means Clustering

In microarray gene expression data analysis, the Euclidean distance and the Pearson correlation coefficient are the two most commonly used similarity measures between genes. Both are tested in the k-means clustering algorithm on the GLIOMA and CAR datasets. The quality of each similarity measure is determined by the average 5-fold classification accuracies for CGS-Cho's/F-test/GS-KNN/SVM methods, under the default setting ($C = 1$ and $k = 100$), i.e., a total of $100 \times 3 \times 2 = 600$ classification accuracies. These averages are plotted in

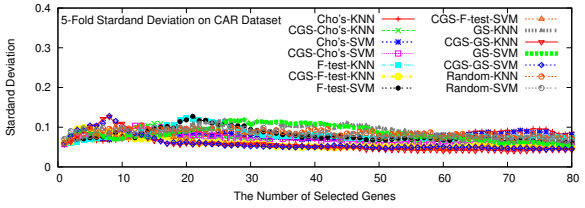

Fig. 4

STANDARD DEVIATION OVER 100 5-FOLD CLASSIFICATION ACCURACIES OF ALL THE SEVEN GENE SELECTION METHODS, COMBINED WITH KNN-CLASSIFIER AND SVM-CLASSIFIER, ON THE CAR DATASET.

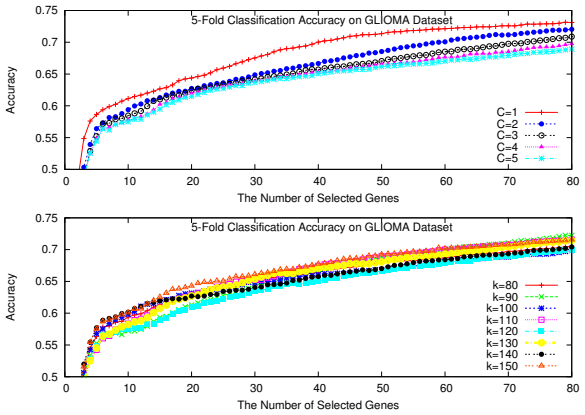

Fig. 5

QUALITIES OF DIFFERENT VALUES OF k AND C, EVALUATED ON THE GLIOMA DATASET. FOR A VALUE OF C, ITS QUALITY IS THE AVERAGE CLASSIFICATION ACCURACY OVER 4800 TESTS; FOR A VALUE OF k, ITS QUALITY IS THE AVERAGE CLASSIFICATION ACCURACY OVER 3000 ONES.

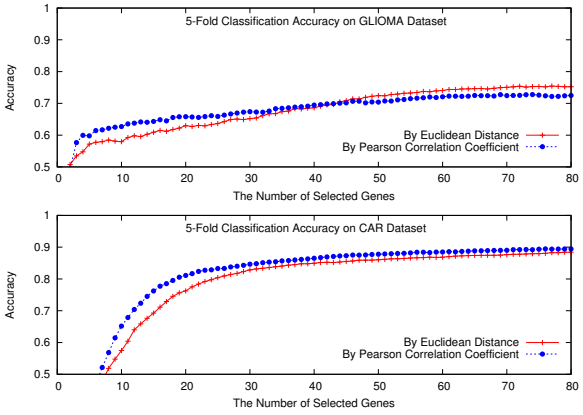

Fig. 6

QUALITY OF THE EUCLIDEAN DISTANCE AND THE PEARSON CORRELATION COEFFICIENT, TESTED ON THE GLIOMA DATASET AND THE CAR DATASET, WHERE $k = 100$ AND $C = 1$.

Figure 6, where one can see that on the CAR dataset, the Pearson correlation coefficient might be (consistently) a little better than the Euclidean distance. But on the GLIOMA dataset such an observation is not noticeable. Overall, the test results on all the eight datasets do not clearly show that one similarity measure is dominantly better than the other. We chose the Pearson correlation coefficient as our default setting. All the results on the Euclidean distance are included in Supplementary Materials.

C. The Performance of CGS

In terms of 5-fold cross validation classification accuracy, the CGS based gene selection methods clearly outperform their non-CGS counterparts (Figures 1 and 2, and Supplementary Materials). To demonstrate the strength of the CGS method more clearly, we took the average of the classification accuracies of all the three CGS based gene selection methods combined with each of the two classifiers, on all the eight datasets (that is, average over $100 \times 3 \times 8 = 2400$ accuracies), and the average of the classification accuracies of all the three non-CGS based gene selection methods combined with each of the two classifiers, on all the eight datasets (that is, again, average over 2400 accuracies). Similarly, for Random, we have Random-KNN and Random-SVM. These six average classification accuracies, CGS-KNN, non-CGS-KNN, CGS-SVM, non-CGS-SVM, Random-KNN and Random-SVM, are plotted in Figure 7, where the averages over LOO cross validation classification accuracies are also separately plotted.

From these 12 plots, one can see that combining with either of KNN-classifier and SVM-classifier, the CGS method can improve the classification accuracy noticeably and consistently, typically when the pre-specified number of selected genes is small.

D. A Case Study: The CAR Dataset

As shown in Figures 1 and 2 and discussed in Section IV.C, the CGS based classification methods are generally superior to the corresponding non-CGS based ones. To determine whether the CGS based methods have chosen biologically relevant genes as the discriminators, the top 80 genes were analyzed further for all six methods CGS-Cho's, Cho's, CGS-F-test, F-test, CGS-GS, and GS. Compared with the tumor-specific 216 genes published in the paper of the CAR dataset [10], the top 80 genes selected by the three non-CGS based methods share 39, 41, 43, respectively (Figure 8), which are roughly 50%. In contrast, the three corresponding CGS methods contain only 17, 17, 20 genes, respectively, which are roughly 25%. The first impression is that this reduced percentage of discriminatory genes doesn't make sense, but further investigation reveals that the CGS methods identify many more cancer related genes, to be explained in the following.

Overall, there are only 3 common genes selected by all the six methods, which are KLK3, ELA3A, and GATA3. KLK3 and ELA3A are known signature genes for prostate and pancreas tumors, respectively [10]. The literature shows that GATA3 is a cancer-related gene too, with its information detailed in Supplementary Materials. This result clearly demonstrates that different gene selection algorithms can pick

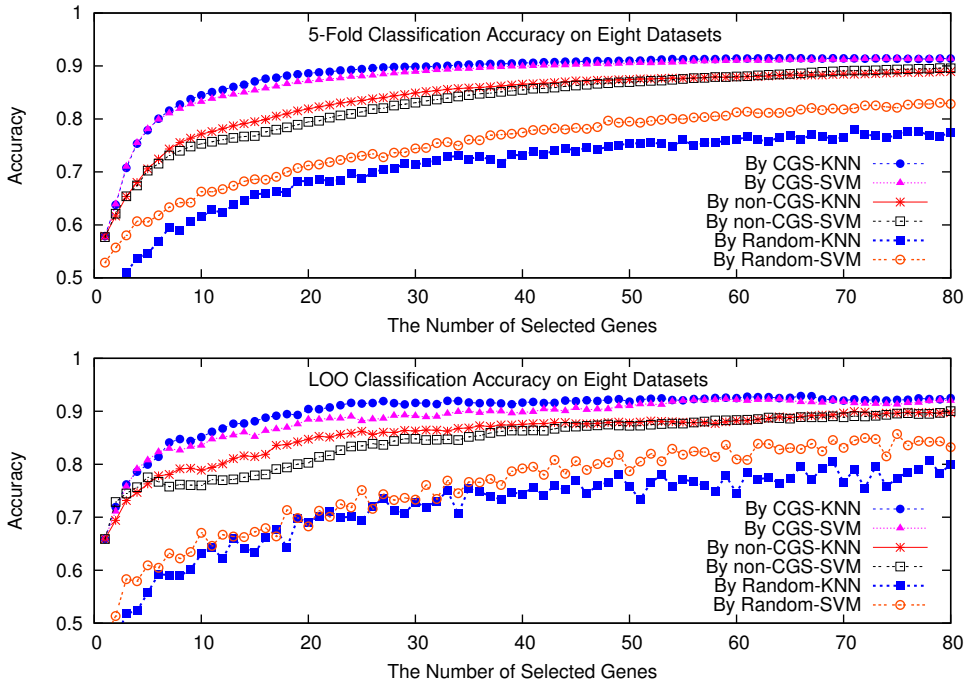

Fig. 7

AVERAGE CLASSIFICATION ACCURACIES OF THE CGS, NON-CGS AND RANDOM BASED GENE SELECTION METHODS, COMBINED WITH KNN-CLASSIFIER AND SVM-CLASSIFIER, IN 5-FOLD AND LOO CROSS VALIDATIONS, ON ALL THE EIGHT DATASETS, RESPECTIVELY.

up different sets of genes for disease classification. The experiments also show that, in principle, *there exist genes having equal discrimination power*, as different subsets of genes can reach quite compatible classification accuracies on the same dataset.

Secondly, it is particularly interesting to find that the gene subsets selected by all three CGS based methods share less with the published tumor-specific genes, but they all achieved notably better classification accuracies. The comparison of these three sets of 80 genes revealed a subset of 29 common genes, which were picked rarely by the three non-CGS based methods (only 3, 9, 10, by Cho's, F-test, GS, respectively). Among these 29 genes, 5 are signature genes published in the original paper, and the other 24 are new discriminatory genes determined by all three CGS based methods. (Note that these 24 probe subsets, see Supplementary Materials, actually represent 23 genes, due to the redundancy of two probe subsets targeting PRCKI.) An immediate question is whether these 24 genes play an important role for achieving improved classification accuracies by the CGS based methods. To answer this question, the probe sets of these 24 new discriminatory genes were examined individually to investigate their biological functions related to the intrinsic molecular characteristics of the cancer cells. The collected facts (see Supplementary Materials) clearly show that the majority of these 24 genes are related to human tumorigenesis, even the probe set *329_s_at*, which does not have any annotations at NETAFFX (Affymetrix database), is decoded to be linked to cancer. Interestingly, these 24 genes were not selected as biomarkers in the original paper [10] and are rarely chosen as top 80 genes in the non-CGS based methods (Figure 8).

Since the classification accuracies of the CGS based methods are always higher than the non-CGS based methods, we believe that the 29 genes (24 new genes and 5 signature genes) shared by the CGS based methods must very likely be the major contributors to the classification improvement. Table I collects the classification accuracy of these 29 genes alone and the classification accuracies of the top 29 genes selected by the three CGS based methods individually, where the former is much higher than all the latter three. We have also tested the six non-CGS based classifiers by using their 33 common genes as the base set of genes, and to include L randomly selected genes, for $L \in [1, 50]$; Similarly, we have tested the six CGS based classifiers by using their 29 common genes as the base set of genes, and to include L randomly selected genes, for $L \in [1, 50]$. The average classification accuracies (each over three) of non-CGS/CGS KNN/SVM classifiers are plotted in Figure 9, where we can conclude that the discrimination power of the 29 CGS common genes is much higher than that of the 33 non-CGS common genes.

Note that introducing CGS to the classification algorithms suggests that the genes with equal discriminatory power might be grouped together based on their similar expression profiles. By limiting the number of genes per cluster, other biologically relevant genes with relatively low ranks can emerge and participate in the list of biomarkers. These new genes, with

TABLE I

CLASSIFICATION ACCURACIES OF THE 33 COMMON GENES AMONG THE TOP 80 GENES SELECTED BY THE THREE NON-CGS BASED METHODS, THE 29 COMMON GENES AMONG THE TOP 80 GENES SELECTED BY THE THREE CGS BASED METHODS, AND THE TOP 29 GENES SELECTED BY THE THREE CGS BASED METHODS, RESPECTIVELY.

	33 non-CGS	29 CGS	CGS-Cho's/29	CGS-F-test/29	CGS-GS/29
5-Fold-KNN	0.6917	0.8931	0.8348	0.8477	0.8612
5-Fold-SVM	0.6330	0.9057	0.8290	0.8417	0.8399
LOO-KNN	0.6954	0.8851	0.8391	0.8506	0.8793
LOO-SVM	0.6264	0.9138	0.8448	0.8333	0.8391

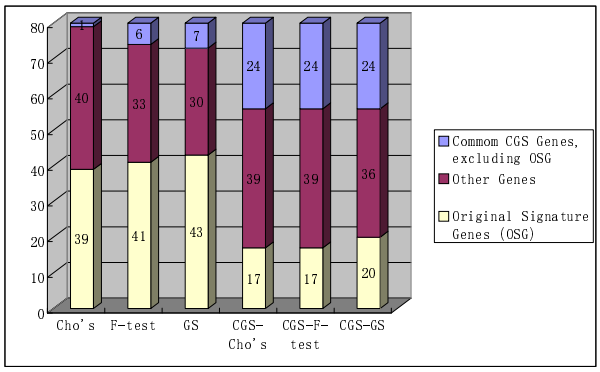

Fig. 8

THE OVERLAPPING GENES AMONG THE TOP 80 GENES RETURNED BY THE SIX METHODS. COMPARED WITH THE TUMOR-SPECIFIC PUBLISHED SIGNATURE GENES, THE SETS OF TOP 80 GENES SELECTED BY THE THREE NON-CGS BASED METHODS CONTAIN 39, 41, 43 SUCH GENES, RESPECTIVELY, SHOWN AS THE BOTTOM PORTION OF THE BARS. THERE ARE 24 NEW COMMON DISCRIMINATORY GENES SELECTED BY THE THREE CGS BASED METHODS, SHOWN AS THE TOP PORTION OF THE BARS.

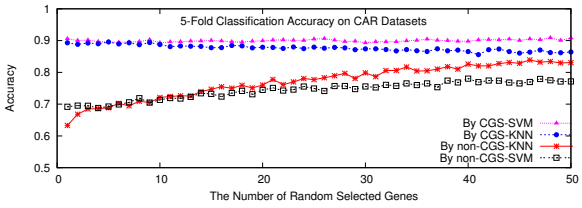

Fig. 9

THE 5-FOLD AVERAGE CLASSIFICATION ACCURACIES OF NON-CGS/CGS KNN/SVM CLASSIFIERS (OVER THREE GENE SELECTION METHODS) USING THE 33 NON-CGS COMMON GENES OR THE 29 CGS COMMON GENES AS THE BASE SET, RESPECTIVELY, AND A NUMBER OF RANDOMLY SELECTED GENES, ON THE CAR DATASET.

different expression patterns, will contribute different strength to classification, to achieve better classification accuracy. The appearance of these 29 genes in the CAR dataset by the CGS based methods has confirmed our initial hypothesis, and shown that the discriminatory genes identified by CGS have an impressively high probability to be tumor-related. Nonetheless, as the parameters C and k were selected based on the test on the GLIOMA dataset, they may not be optimal with respect to the CAR dataset. For example, we still found redundant biomarkers by the CGS based methods, as well as two probe subsets targeting the same gene PRCK1 in the 29 gene set. One possible resolution of this issue is to increase the k value. It is very likely that 100 clusters is too small for the CAR dataset. However, the highly improved classification results with the non-optimized parameters indicated that the values of these parameters are very robust in the CGS based methods, though better classification accuracy could still be achieved by fine tuning of these parameters.

For the collection of 500 subsets of 80 genes, each from a 5-fold cross validation (100 independent runs) for CGS-Cho's-KNN on the CAR dataset, we calculated the frequency of a gene occurring in these 500 subsets. It turns out that there are 14 genes whose frequencies go beyond 90% (Figure 10), indicating that the selected genes are quite stable by the CGS-GS-KNN method. Moreover, the average frequency of the 29 common genes by the CGS methods is 0.7587, which is much higher than the average frequency of other genes, 0.5421. It provides another evidence that these 29 common genes could be biologically more meaningful than the others.

V. CONCLUSIONS

We have examined a method of combining gene clustering to identify many biological relevant discriminatory genes for class prediction using microarray datasets. Our clustered gene selection (CGS) for classification is able to identify certain clusters of genes that have equal strength in classification, yet is computationally very efficient. Our experiments on the eight real datasets showed that CGS can clearly improve the classification accuracy by selecting the same number of genes, compared to three non-CGS based gene selection methods. Moreover, there are many interesting properties associated with the CGS method, for example, the CGS based gene selection methods tend to have much bigger set of overlapping genes.

The three non-CGS based gene selection methods Cho's, F-test, and GS all use correlation coefficients to rank the genes. Our next immediate task is to examine the performance of the CGS method when combined with other categories of gene selection methods, such as using the individual gene contribution to the separation as a ranking scheme [6] as well as the relatively more complex SVM-RFE gene selection method by Guyon *et al.* [7].

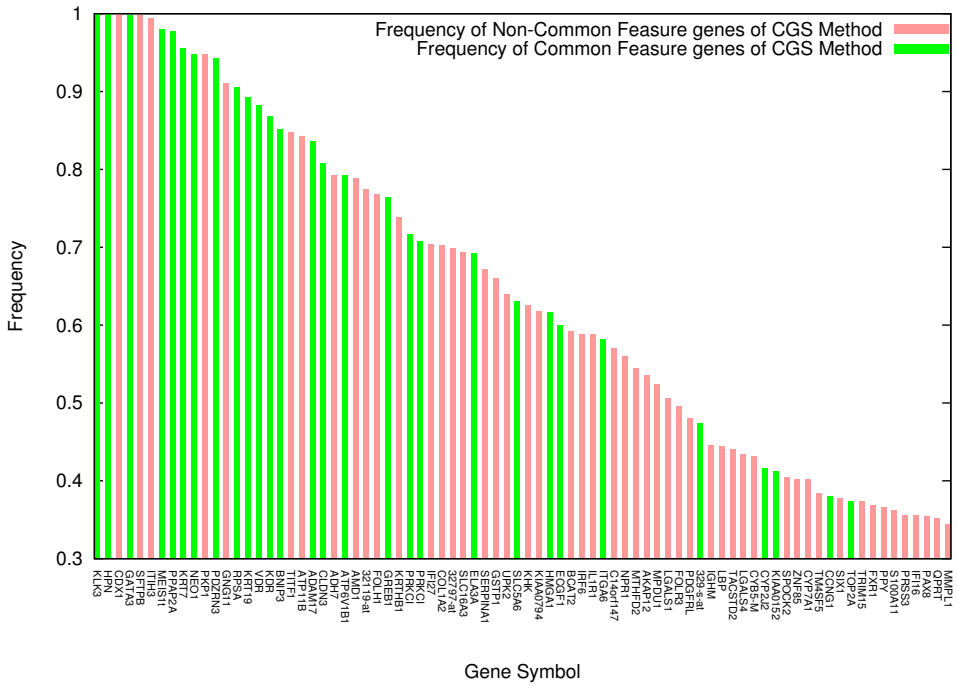

Fig. 10

THE FIRST 80 GENES, SORTED IN DECREASING FREQUENCY, SELECTED BY THE CGS-CHO'S-KNN METHOD ON THE CAR DATASET, COLLECTED IN 500 SUBSETS OF 80 GENES EACH IN THE 5-FOLD CROSS VALIDATION. THE 24 COMMON GENES SELECTED THE THREE CGS BASED METHODS ARE SHOWN IN SOLID GREEN BARS.

VI. SUPPLEMENTARY MATERIALS

The eight microarray datasets are provided through webpage "`http://www.cs.ualberta.ca/~ghlin/src/WebTools/cgs.php`". In the same webpage, many more experimental results are included, which could not sit here due to space limit.

ACKNOWLEDGMENTS

LX's research was done while visiting the University of Alberta and partially supported by AHFMR. ZC, YS, MRS, RG and GL are grateful to the research support from AICML, CFI, NSERC and the University of Alberta. All authors would like to thank the two reviewers for their many constructive comments on the submission version of this paper.

REFERENCES

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[2] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[3] K. Yang, Z. Cai, J. Li, and G.-H. Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7:228, 2006.

[4] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[5] J. Cho, D. Lee, J. H. Park, and I. B. Lee. New gene selection for classification of cancer subtype considering within-class variation. *FEBS Letters*, 551:3–7, 2003.

[6] M. Xiong, X. Fang, and J. Zhao. Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887, 2001.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[8] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–14868, 1998.

[9] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. V. Deimling, S. L. Pomeroy, T. R. Golub, and D. N. Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63:1602–1607, 2003.

[10] A. I. Su, J. B. Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk, H. F. Frierson, Jr., and G. M. Hampton. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*, 61:7388–7393, 2001.