World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:2, 2010

Novel Hybrid Method for Gene Selection and Cancer Prediction

Liping Jing, Michael K. Ng, and Tieyong Zeng

Abstract—Microarray data profiles gene expression on a whole genome scale, therefore, it provides a good way to study associations between gene expression and occurrence or progression of cancer. More and more researchers realized that microarray data is helpful to predict cancer sample. However, the high dimension of gene expressions is much larger than the sample size, which makes this task very difficult. Therefore, how to identify the significant genes causing cancer becomes emergency and also a hot and hard research topic. Many feature selection algorithms have been proposed in the past focusing on improving cancer predictive accuracy at the expense of ignoring the correlations between the features. In this work, a novel framework (named by SGS) is presented for stable gene selection and efficient cancer prediction . The proposed framework first performs clustering algorithm to find the gene groups where genes in each group have higher correlation coefficient, and then selects the significant genes in each group with Bayesian Lasso and important gene groups with group Lasso, and finally builds prediction model based on the shrinkage gene space with efficient classification algorithm (such as, SVM, 1NN, Regression and etc.). Experiment results on real world data show that the proposed framework often outperforms the existing feature selection and prediction methods, say SAM, IG and Lasso-type prediction model.

Keywords—Gene Selection, Cancer Prediction, Lasso, Clustering, Classification.

I. Introduction

CANCER is a class of diseases for which a group of cells undergoes uncontrolled growth. It causes destruction of adjacent tissues and sometimes spreads to other locations in the body via lymph or blood. American Cancer Society stated that about 7.6 million people died from cancer in the world during 2007, and nearly all cancers are caused by abnormalities in the genetic material of the transformed cells. Various research efforts based on surgery, chemotherapy and radiotherapy, are being made to fight against cancer. Recently, more and more researchers began to study gene expression profiles obtained by microarray technology. Microarray data profiles gene expression on a whole genome scale and provides a good way to study associations between gene expression and occurrence or progression of cancer [2]. It has been used extensively in variety of applications, ranging from basic molecular biology research, through testing drug treatment effectiveness, and up to clinical diagnosis of cancer patients based on their gene expression profiles. Thus, microarray data analysis has a profound impact on cancer research [1].

In microarray data analysis, there is a big challenging problem, the dimension of gene expressions is much larger than the sample size, which makes it be a hot and hard research topic [23], [24]. In order to solve this problem, feature selection [3], a technique of selecting a minimum subset of original features for best predictive accuracy, has attracted strong interest in the past several decades for text mining, image processing and etc.. Among them, researchers made use of feature redundancy techniques [4] which minimize redundancy and maximize relevance among selected features for classification. In the field of bioinformatics, a large amount of efforts have also been made to identify relevant or important genes that have influential effects on diseases including varieties of cancers. For example, statistical approaches for gene selection and predictive model building have been widely studied and designed. Previously employed approaches include the singular value decomposition [5], principal component analysis [6], [7], partial least squares [8], sparse logistic regression [12], [13], Lasso [9], [17], [46], support vector machine [10], [11], information gain [28], fuzzy theory [25], SAM [31] and etc. These approaches aim at improving the cancer prediction accuracy by identifying the individual genes, a small subset of genes or linear combinations of genes often referred as super genes which can best explain the phenotype variations.

In real cancer data, many different subsets of genes may result in the same or similarly good sample class prediction accuracy [14]. However, the above methods are not necessarily reliable to identify such candidate genes for subsequent costly biological validation, even though they are effective in cancer sample prediction. There are two reasons causing such situation. One reason is the classic goal of gene selection methods, which discards the genes relevant to the target concept but highly correlated to the selected genes. Therefore, among a set of highly correlated genes, different genes may be selected under different settings of a selection algorithm. Recent work has confirmed that the feature selection algorithms can obtain different subsets of features under training data variations [15]. The other reason is the relatively small number of samples in high-dimensional data.

In order to handle the above problem, Unger and Chor [16] presented a linear separating method (named as LinSep) to find all gene pairs such that the projection of all samples according to each gene pair can be separated, which to the large extent

Liping Jing is with School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, P.R.China.Email: lpjinghk@gmail.com.

Michael K. Ng and Tieyong Zeng are with Department of Mathematics, Hong Kong Baptist University, Kowloon Toog, Hong Kong. Email: mng@math.hkbu.edu.hk, zeng@hkbu.edu.hk.

Tieyong Zeng is the Corresponding Author.

Part of research was supported by Hong Kong RGC (201508), HKBU FRGs, the Research Fund for the Doctoral Program of Higher Education (200802691037) and the National Natural Science Foundation of China (90820013, 60875031,60905028), 973 project (2007CB311002).

identified all important genes for sample class prediction and meanwhile considered the relationship between genes. LinSep has to project the samples on all gene pairs and then select the best gene pairs, thus it is computationally expensive. Yu et al. [32] and Loscalzo et al. [33] proposed stable feature selection approaches via dense and consensus feature groups, DRAGS and CGS respectively. They firstly identify feature groups where all features in each group are as much correlated to each other as possible, and then apply the selection methods on the feature group level where each group is treated as a single entity. However, it is better that DRAGS and CGS used more samples to effectively identify the feature groups firstly, meanwhile, which adds the uncertainty of the final feature selection results. Yuan and Lin [21] presented group Lasso (named as grpLasso) to identify the important gene groups where the covariates are partitioned into groups. DRAGS, CGS, grpLasso are able to find the important correlated gene groups but they can not identify the important individual genes in each gene group, i.e., they can not automatically determine the group size.

In this paper, a hybrid strategy was presented to effectively select important genes from microarray data and accurately classify the cancer samples. The proposed method combines Clustering approach, the Bayesian Lasso approach, the group Lasso approach and Classification approach. These four approaches have essential connections. Here, the clustering approach provides the gene group label based on genes' correlation coefficients to the later group Lasso, the Bayesian Lasso method identifies a stable subset of important genes for each group to re-represent the cancer data in the shrinkage space, the group Lasso selects the important gene groups based on the obtained group labels and the shrinkage data representation, and finally the classification approach builds a classifying model to predict cancer data based on the important shrinkage space. As these methods are related to each other, their integration is consistence and thus expected to provide efficient results. A series of experimental results have shown that the proposed strategy performs well in real applications. The rest of our paper is organized as follows. In Section 2, some related and typical gene selection methods in bioinformatics will be given. Section 3 will describe the proposed hybrid framework. Section 4, a series of experimental results will be shown and discussed. A conclusion will be given in Section 5.

II. Related Works

Gene selection is necessary, important and difficulty for cancer identification [23]–[25], therefore, is a hot research topic. So far, there are many available statistical approaches for gene selection and predictive model building which consider the property of microarray data, the dimension is much larger than the sample size. In this section, several popular methods for gene selection in microarray data analysis will be briefly reviewed, such as Information gain [28], SAM mehtod [31] and Lasso-type methods [17], [18], [21], [46].

Information gain (IG) is a feature-goodness criterion in the field of machine learning [29]. Recently, it was used in text mining [26], [27], and predictive gene identification [28]. IG measures the amount of information that presence or absence of a particular gene contains about the category of the sample.

$$IG(i) = \sum_{c\in\{c_1,c_2\}} \sum_{g\in\{g_i,\bar{g_i}\}} P(g,c) \times log \frac{P(g,c)}{P(g) \times P(c)} \quad (1)$$

where $P(g)$, $P(c)$, $P(g,c)$ measure the probability of a gene, a tissue category (cancer or normal), and both a gene and a tissue category appear in the whole cancer sample set respectively. The IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a gene in a sample.

Tusher et al. proposed a method to analyze the significance of genes for ionizing radiation response (SAM method [31]). The relative difference of each gene is defined as:

$$SAM(i) = \frac{\mu_1(i) - \mu_2(i)}{s(i) + s_0} \quad (2)$$

$$s(i) = \sqrt{\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2} \{\sum_{j=1}^{n_1} [x_{1j}(i) - \mu_1(i)]^2 + \sum_{j=1}^{n_2} [x_{2j}(i) - \mu_2(i)]^2\}}$$

where $\mu_1(i)$ and $\mu_2(i)$ are the average levels of expression for the ith gene in class 1 and 2 respectively. $s(i)$ is the standard deviation of repeated expression measurements (or the samples), called as gene-specific scatter. n_1 and n_2 are the numbers of samples in class 1 and 2.

The above methods, IG and SAM, select genes according to their corresponding score. The higher score, the more important genes. In order to identify important genes, these methods require a threshold to cut the whole gene sets and then keep the genes with larger score. However, in real application, it is hard to choice an appropriate threshold.

Recently, the least absolute shrinkage and selection operator (Lasso), a shrinkage method, has been widely used in regression analysis for large models [17]. The Lasso procedure can be interpreted as a Bayesian posterior mode estimate when assigning an independent double-exponential prior to each coefficient [17], [21], [46]. Owing to the nature of the L_1-penalty, the lasso does both continuous shrinkage and automatic variable selection simultaneously. Standard Lasso approach carries out variable selection at the individual gene level and takes the form

$$Lasso(\lambda, \beta) = ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|_1. \quad (3)$$

Although the Lasso penalty leads to sparse models, it does have two serious drawbacks. Firstly, Lasso is instable when the data is high-dimensional. In order to deal with this problem, several researchers proposed new Lasso-type estimators based on Tibshirani' analysis result [17] that Lasso can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors. Zou [19] proposed an adaptive Lasso estimator by introducing adaptive data-driven weights and Laplace-like priors. Park and Casella [46] adopted Bayesian posterior mode estimate and marginal maximum likelihood to automatically find the best parameter. Zou and Hastie

[18] proposed an elastic net estimator (enLasso) to select the correlated variables as follows.

$$enLasso(\lambda, \lambda_1, \beta) = ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|_1 + \lambda_1 \sum_{j=1}^{p} ||\beta_j||_2^2 \quad (4)$$

The l_1-norm part of Eq.(4) performs automatic variable selection, while the l_2-norm part stabilizes the solution paths and, hence, improves the prediction. However, parameter λ_1 for l_2-norm penalty makes elastic net become hard control, so does adaptive elastic net estimator [20].

Secondly, Lasso does not contain any prior information about, e.g., possible groups of covariates that one may wish to select them jointly. Several researchers have recently proposed new penalties to enforce the estimation of models with specific sparsity patterns. Yuan and Lin [21] presented group lasso to deal with the case where the covariates are partitioned into groups. The group lasso [21], [22] (referred as grpLasso) is designed for selecting groups of covariates by optimizing the following estimator

$$grpLasso(\lambda, \beta) = ||y - X\beta||_2^2 + \lambda \sum_{g=1}^{G} ||\beta_{I_g}||_2, \quad (5)$$

where I_g is the index set belonging to the gth group of genes, $g = 1, \ldots, G$. This penalty can be viewed as an intermediate between l_1 and l_2 type penalty. It has the attractive property that it can select gene at the group level and is invariant under (groupwise) orthogonal transformations like ridge regression. Direct application of the grpLasso can identify important gene groups. However, it is not capable of selecting important genes within the selected groups.

III. Novel Framework for Gene Selection and Cancer Prediction

In this section, a novel framework for both selecting important genes and building cancer prediction model will be presented. The proposed framework effectively combined several data mining techniques: clustering algorithms based on Pearson coefficient, sparse feature selection methods, covariates selection methods and classification algorithms, as shown in Fig.1. The new framework adopted the combination of these four techniques, so that kept the merits of existing algorithms (such as, BLasso for sparse individual feature selection and Group Lasso for covariates identification) and filled up the drawbacks of them (say, BLasso can not identify covariates, while Group Lasso can not select important genes in each group [45]).

In the proposed framework, partitioning around medoids (PAM) [42] clustering algorithm was adopted to identify the gene groups. PAM represent each cluster with one of its own object, i.e., the representative object is a centrotype. PAM is more robust than kMeans because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances, which is one reason PAM algorithm was used here. The other reason is that PAM operates on the dissimilarity matrix of the given cancer gene expression data, in this case, the dissimilarity between two genes (x and y) is calculated with

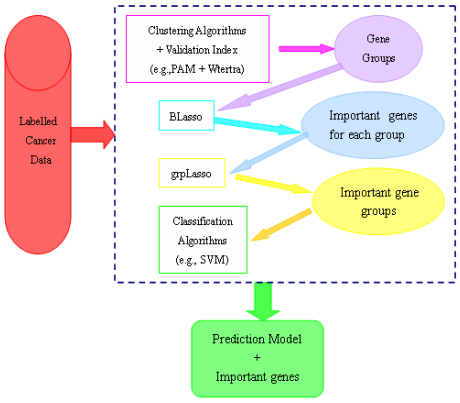

Fig. 1. The framework for both important gene selection and prediction model building

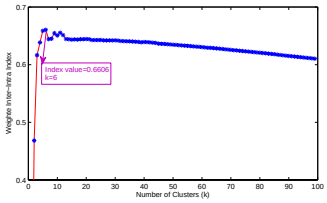

Fig. 2. Clustering validity index (Weighted inter-intra index) as function of number of clusters for a) Colon data set and b) Leukaemia data set

the Pearson Correlation Coefficients (as shown in Eq.(6) which is a popular and efficient similarity metric in gene expression profiles analysis [44])

$$r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \quad (6)$$

The dissimilarity between x and y is defined by $d(x, y) = 1 - \frac{1+r(x,y)}{2}$, where n is the number of samples in the cancer data. Meanwhile, Pearson coefficient will be helpful for later grpLasso step of our framework.

However, like kMeans, PAM required the number of clusters k as the input parameter. Here, a big range of k was tested and the best k was identified with weighted inter-intra index ($Wtertra$) [43].

$$Wtertra = (1 - \frac{2k}{p})(1 - \frac{\sum_{i=1}^{k} \frac{p_i}{p - p_i} \sum_{j=1}^{k} p_j inter(i,j)}{\sum_{i=1}^{k} p_i intra(i)} \quad (7)$$

where

$$inter(i, j) = \frac{1}{p_i p_j} \sum_{x \in C_i, y \in C_j} Similarity(x, y)$$

and

$$intra(i) = \frac{2}{p_i(p_i - 1)} \sum_{x,y \in C_i, x \neq y} Similarity(x, y)$$

TABLE I
Gene Clusters Size of Colon Data Set

ClusterID	1	2	3	4	5	6
♯Genes	227	369	221	331	238	614

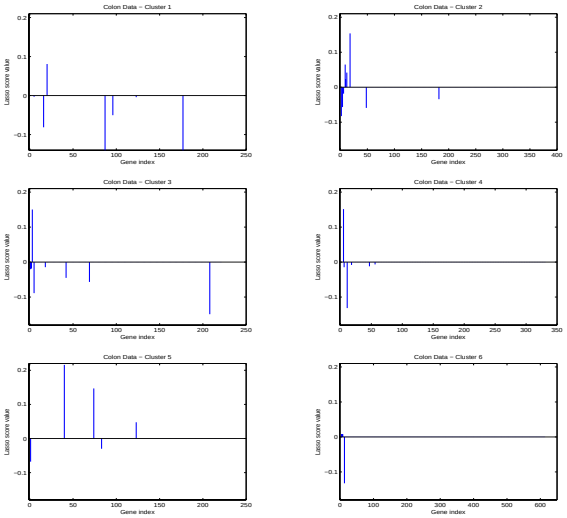

Fig. 3. Feature weights obtained by Lasso for each gene group cluster on Colon Data Set.

p is the number of total genes, p_i is the number of genes in the ith cluster, $inter(i,j)$ is the similarity between the ith and the jth clusters, while, $intra(i)$ is the similarity in the ith cluster. In order to find a desirable clustering result, with high overall cluster quality (i.e., maximizing intra-cluster similarity and minimizing inter-cluster similarity) and a small number of clusters k, the clustering result with the highest $Wtertra$ value was adopted. In other words, PAM with Pearson Correlation Coefficient and $Wtertra$ index can identify the gene groups where the genes in each group are covariates because they have higher correlation coefficient (i.e., similarity).

Let us take a real cancer data set, Colon data including 2000 genes and 62 samples [34], as an example to show the performance of PAM and $Wtertra$. Fig.2 gives the clustering validation results (k from 2 to 100) of PAM on Colon data set. The best number of gene groups in Colon data is 6 according to Fig.2. Table I gives the cluster sizes at the best clustering result on Colon data set. The clustering step provides the gene group label to supervise the grpLasso step in our proposed framework.

So far, group Lasso [21] can be used to identify the important gene clusters for cancer prediction with the above obtained gene cluster labels, but grpLasso has bad ability to select the important individual genes [45]. Therefore, Lasso [17], a technique encouraging sparsity in individual coefficients based on a small set of samples, was applied on each gene group to identify a small subset of informative genes, called marker genes, which discriminate between the tumor and the normal tissues, or between different kinds of tumor tissues. Recall the above Colon example, six gene groups were obtained, which means that Lasso would be applied six times and once for each group. Considering the difficulty of parameter selecting in Lasso, here, Bayesian Lasso (BLasso) [46] was adopted which can approximate the ideal updated penalty parameter λ with marginal maximum likelihood during the iteration process. Fig.3 lists the gene (variable) coefficients distribution for each Colon gene group, where the genes with non-zero coefficient are marker genes, i.e., important to discriminate the cancer and normal tissue samples.

Once the important genes are identified in each group, all tissue samples can be represented with these important genes and their group labels. For example, the genes are clustered into K groups by PAM and $Wtertra$ index, and the kth group has m_k important genes identified by BLasso, then the tissue sample X_i will be represented as a vector with $D = \sum_{k=1}^{K} m_k$ dimensions like $X_i = \{f_{k1}, f_{k2}, \cdots, f_{km_k} | 1 \leq k \leq K\}$. Finally, the cancer data set is re-represented as a n-by-D matrix. In the new representation model, the genes in each group are more similar than the genes in different group. Among them, the similarity between genes are calculated by the Pearson correlation coefficient metric, therefore, genes in the same group are covariates in the whole data set. Recall the theory of group Lasso (Eq.(5)), identifying the group of covariates, grpLasso can be applied on the new cancer data representation under the supervision of group labels.

As we know, the accuracy of the classification model depends strongly on how the input data is represented. Typically, the input data is transformed into a feature vector containing a number of features that are descriptive of the data. Because of the curse of dimensionality, the number of features should not be too large, but should be large enough to accurately predict the output. In our proposed hybrid framework for cancer prediction, the number of final selected important genes (i.e., features) are not too large but large enough to build the classification model, because these genes are identified by an effective integrated method which combines clustering, Bayesian Lasso and group Lasso. For the classification models, support vector machine (SVM) [47] k-nearest neighbors (KNN) [48] and logistic regression [49] were adopted. SVM is a popular and efficient linear classifier by finding a hyperplane so that the distance from it to the nearest data points on each side is maximized, especially, for two-class cancer prediction case. KNN algorithm is the simplest machine learning algorithm which classifies a data point by a majority vote of its neighbors, in our case, K was set to be 1, i.e., the data is simply assigned to the class of its nearest neighbor. Also, logistic regression can be used here to predict the probability of occurrence of an event by fitting data to a logistic curve. Actually, our framework can adopt any classification algorithm in this step.

In the next Section, the proposed method will be applied on the real cancer data sets and the experimental results will be reported. The proposed method combines Clustering approach, the Bayesian Lasso approach, the group Lasso approach and Classification approach. These four approaches have important connections. Indeed, the clustering approach provides the gene group label based on their correlation coefficients to the later group Lasso, the Bayesian Lasso method identifies a series of important genes for each group to re-represent the cancer data in the shrinkage space, the group Lasso selects the important gene groups based on the obtained group labels and the shrinkage data representation, and finally the classification approach builds a classifying model to predict cancer data based on the important shrinkage space. The clustering algorithm measures the correlation between features effectively, thus it

is helpful for the group Lasso to select the import covariates. The Bayesian Lasso effectively shrinks the sparsity of the gene space, therefore, it is useful for the group Lasso and classification model building. As these methods are related to each other, their integration is consistence and is thus expected to provide efficient results.

IV. Experimental Results and Discussion

The proposed hybrid framework was tested with the real cancer data, Colon cancer data [34], Leukaemia data [30] and breaset cancer data [35]. The colon cancer data describes the expression of 2000 genes in 40 cancer and 22 normal tissue samples, the aim being to construct a classifier capable of distinguishing between cancer and normal tissues. The aim of the leukaemia benchmark is to form a decision rule capable of distinguishing between acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL). The data describes the expression of 7129 genes in 47 ALL samples and 25 AML samples. The breast cancer data set is consisted of 49 tumour samples in 7129 genes. This data provides two functions. One is to build a classifier to distinguish into estrogen receptor-positive (ER+) tumour samples and estrogen receptor-negative (ER-) ones, the other is to classify the tumor samples lymph node-positive (LN+) and lymph node-negative (LN-).

For SAM and IG scoring methods, the number of genes was selected same with the number of genes obtained by SGS method, and applied SVM and 1NN classification algorithms on the selected gene space. BLasso and grpLasso were compared with the proposed framework in two ways. One is using BLasso or grpLasso as the gene selection method and then adopt SVM and 1NN to build the cancer prediction model. For the whole data set, the best penalty parameter λ of BLasso was automatically determined during the iteration, and the corresponding numbers of selected genes are 27 for Colon, 71 for Leukaemia, 48 for Breast (LN) and 48 for Breast (ER) respectively. The other way is directly running BLasso and grpLasso to predict the samples with logistic regression method, and their results were compared with the prediction results obtained by the integrated method SGS and logistic regression model.

Our proposed hybrid method identified that the best cluster numbers are 6 for Colon, 54 for Leukaemia, 49 for Breast (LN) and 54 for Breast (ER) respectively (for all data, set k from 2 to 100). The number of selected genes are 20 for Colon, 49 for Leukaemia, 33 for Breast (LN), and 36 for Breast (ER). Table II and III show the classification accuracy of the above four data sets, The first evaluation method (Table II) is 10-fold validation on training data, and the second one (Table III) is classification accuracy on test data. According to the experimental results, the proposed hybrid gene selection method (SGS) significantly outperforms the existing feature selection methods BLasso, grpLasso, SAM and IG.

Meanwhile, the gene weights (for Leukemia data) and their distribution were listed, where gene weights were obtained by these three methods (SGS (Fig.4(a)), SAM (Fig.4(b)), and IG (Fig.4(c)).) and the weights distribution between the proposed SGS and the other two methods SAM (Fig.4(d)) and IG (Fig.4(e)). Fig.4(a) makes us conveniently select the important genes with non-zero weights, however, SAM and IG scoring methods have to predefine a threshold to choose the significant genes. The comparison distributions between SGS and SAM scores, and between SGS and IG scores indicate that not all genes with higher SAM score or IG score have higher SGS score. Meanwhile the selected genes by SGS (as shown in Table V) were checked by the biologists and they found that most of such genes are empirically useful to determine ALL and AML, thus SGS methods are more reasonable than SAM and IG to identify significant genes for cancer prediction. The reason why SGS performs better than BLasso and grpLasso is that SGS efficiently integrated both of them and clustering algorithm, so that SGS keeps the merits of BLasso and grpLasso and simultaneously filled up their drawbacks. Furthermore, comparing with the prediction accuracy 82.3% of BLogReg on Colon data [13], 93.1% of BLogReg on Leukemia data [13], 81.9% of SLogReg on ER Breast data [12], 92.1% of enLasso on Leukemia data [18], our proposed hybrid framework also performs better than them.

Furthermore, the final selected genes were checked, especially for Colon data and Leukemia data, to confirm whether the identified genes are related to the cancer samples.

Table IV shows the selected significant genes from Colon cancer data. Among them, Gene Hsa.3016 is S-100P protein. 100P is well-known expressed in human cancers, including breast, colon, prostate, and lung, therefore, its expression level was correlated with resistance to chemotherapy. Similar to our work, [36] also identified that Hsa.1039 and Hsa.627 were associated with Colon cancers. Hsa.140 and Hsa.1737 were also selected by [37], although the gene selection procedure adopted there was different from the one used in the proposed method. Gene Hsa.462[1] has been shown to be significantly correlated to colon tissues. Bolmont et al. [38] proved that Human desmin gene (Hsa.8147) played an important role in colon disease. Hsa.36689, Hsa.37937 and Hsa.2291 were identified by [12] and the first two and Hsa.6814 were also identified by [39]. Tristetraprolin (Hsa.1682) was examined to have ability to regulate COX-2 which increases in colon tumor microenvironment, that is, Hsa.1682 is observed during colon tumorigenesis [40]. Hsa.8214 has been shown to be associated with tumor cell proliferation in general, while Hsa.696 was revealed to be related to small intestine, colon, testis, and leukocytes by RT-PCR analysis. Meanwhile, researchers revealed that Hsa.1454 acts as a negative regulator of the LEF-1/beta-catenin transcription complex, thereby protecting cells from development of cancer [41]. Estrogen sulfotransferase (Hsa.42949) can inhibit competitively the activation of promutagenic estrogen metabolities into carcinogens, so that it has protective effect for colon cancer.

Table V listed the important 49 genes for Leukemia data identified by the proposed SGS method. Among them, 44 genes have been empirically proved to be discriminative genes between ALL and AML [30]. For the other five genes M27830, M11722, M12886, M14483 and X00437 , although there is no exact biological experiments to show they can determine

[1] http://www.nextbio.com/b/search/ov/SERPINC1?type=feature

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY (% 10-FOLD VALIDATION ON ALL DATA) WITH DIFFERENT METHODS

	SVM					1NN					Regression		
Dataset	SGS	BLasso	grpLasso	SAM	IG	SGS	BLasso	grpLasso	SAM	IG	BLasso	grpLasso	SGS
Colon	90.5	87.1	87.1	88.7	87.1	85.5	88.7	85.5	82.3	85.5	88.7	87.1	83.8
Leukaemia	95.8	98.6	98.6	91.7	93.1	96.7	97.2	97.2	95.8	94.4	94.4	94.4	95.8
Breast (LN)	94.1	49.0	63.3	81.6	77.5	100	73.5	67.3	83.7	85.7	83.7	84.5	91.8
Breast (ER)	91.8	67.3	71.4	77.5	83.7	87.8	81.6	87.8	83.7	85.7	89.8	87.8	91.8

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY (% 10-FOLD VALIDATION ON TESTING DATA) WITH DIFFERENT METHODS

	SVM					1NN					Regression		
Dataset	SGS	BLasso	grpLasso	SAM	IG	SGS	BLasso	grpLasso	SAM	IG	BLasso	grpLasso	SGS
Colon	90.5	90.5	90.5	90.5	85.7	90.5	85.7	90.5	71.4	85.7	90.5	90.5	95.2
Leukaemia	100	100	100	100	95.8	96.7	87.5	95.8	95.8	94.4	95.8	95.8	95.8
Breast (LN)	94.1	58.8	70.6	81.6	77.5	100	64.7	76.5	83.7	85.5	82.4	88.2	88.2
Breast (ER)	94.1	35.3	35.3	76.5	76.5	100	94.1	70.6	82.4	82.4	87.8	76.5	91.8

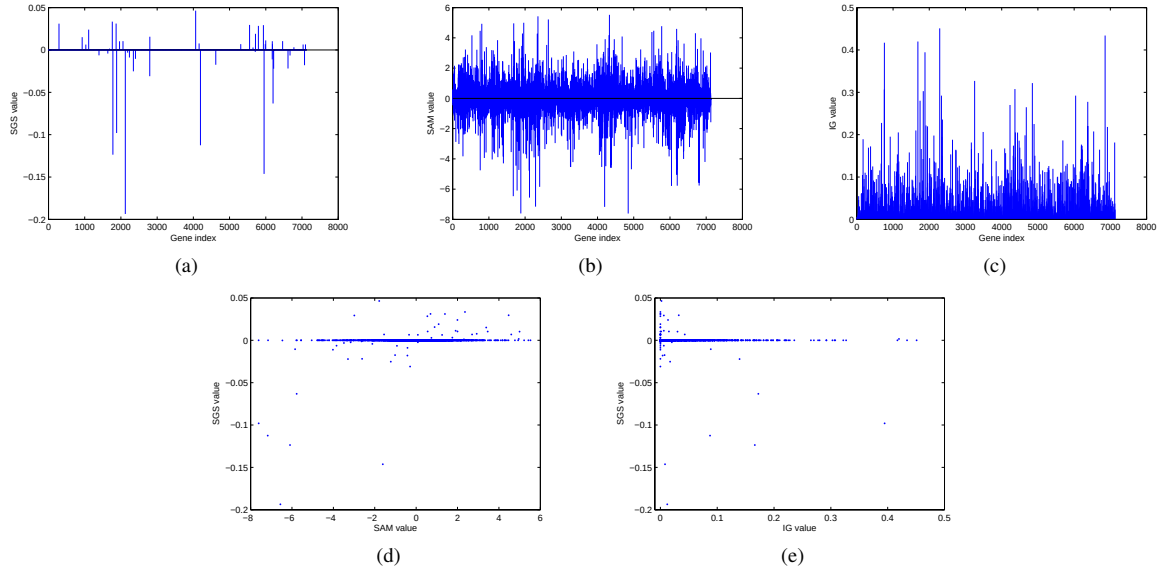

Fig. 4. The gene weights obtained by different methods: a) the proposed SGS method, b) SAM method, c) Information gain method, and the distribution d) between SGS weights and SAM weight, e) between SGS weight and IG weight

the difference between ALL and AML, our results can help the biologist to design future lab experiments to confirm the confidence of the relationship between gene and cancer samples. As we know, when biologists choose gene candidates, due to the difficulties intrinsic in the biological experiments, it is not feasible for them to validate a large number of genes. Therefore, our experimental results are meaningful.

V. CONCLUSIONS

A hybrid gene selection method was presented in our paper. The hybrid method effectively integrated three techniques, clustering, Baybesian Lasso and group Lasso, so that it can identify both the important individual gens and their correlated genes. The clustering algorithm based on Pearson coefficient metric provides supervision information (e.g., group labels) for group Lasso, and Bayesian Lasso extracted important individual genes for each group, thus the final step, group Lasso, can efficiently find the significant gene groups where each group only contains the important correlated genes. Based on this hybrid gene selection method, the original cancer data is represented in the selected genes space, and then any classification algorithm (such as, SVM, 1NN, Regression and etc.) can be applied to build the prediction model. Experimental results on four cancer data sets have shown that our proposed method (SGS) always performs better than the existing gene selection methods, say, SAM, IG, Lasso-type.

REFERENCES

[1] T. Golub: Genome-wide views of cancer. New England Journal of Medicine, 344, 8, 601-602, 2001.
[2] S. Ramaswamy, T. Golub: DNA microarrays in clinical oncology. Journal of clinical oncology, 20, 7, 1932-1941, 2002.
[3] H. Peng, F. Long, C. Ding: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern analysis and machine intelligence, 27, 1226-1238, 2005.
[4] A. Appice, M. Ceci, S. Rawles, P. Flach: Redundant feature elimination for multi-class problems. Proc. of the 21st ICML, 33-40, 2004.
[5] T. Golub, C. van-Loan: Matrix Computations baltimore. Johns Hopkins Univ. Press, 1996.
[6] S. Ma, M. Kosorok, M, J. Fine: Additive risk models for survival data with high dimensional covariates. Biometrics, 62, 202-210, 2006.

TABLE IV
SIGNIFICANT GENES OF COLON DATA SELECTED BY THE PROPOSED SGS METHOD

CI	GeneID	Gene Discription
1	Hsa.3087	TRANSLATIONALLY CONTROLLED TUMOR PROTEIN (HUMAN)
	Hsa.3016	S-100P PROTEIN (HUMAN)
	Hsa.1039	Homo sapiens secretory pancreatic stone protein (PSP-S) mRNA
2	Hsa.140	IG GAMMA-1 CHAIN C REGION (HUMAN)
	Hsa.1737	IG KAPPA CHAIN PRECURSOR V-III REGION (HUMAN)
	Hsa.1272	ATP synthase coupling factor 6, mitochondrial precursor
4	Hsa.467	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN)
	Hsa.4689	40S RIBOSOMAL PROTEIN S18 (Homo sapiens)
	Hsa.462	Human serine kinase mRNA, complete cds
	Hsa.627	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA
5	Hsa.8147	Human desmin gene, complete cds
	Hsa.36689	H.sapiens mRNA for GCAP-II/uroguanylin precursor
	Hsa.2291	GELSOLIN PRECURSOR, PLASMA (HUMAN)
	Hsa.37937	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
	Hsa.1682	TRISTETRAPROLINE (HUMAN)
6	Hsa.6814	Collagen Alpha 2 (XI) Chain (Homo sapiens)
	Hsa.8214	PUTATIVE SERINE/THREONINE-PROTEIN KINASE B0464.5 IN CHROMOSOME III (Caenorhabditis elegans)
	Hsa.696	Human cleavage stimulation factor, complete cds
	Hsa.1454	Human gamma amino butyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds.
	Hsa.42949	ESTROGEN SULFOTRANSFERASE (Bos taurus)

[7] J. Costa, H. Alonso, L. Roque, A weighted principal component analysis and its application to gene expression data, IEEE/ACM Trans. on computational biology and bioinformatics, 17 Jul. 2009. IEEE computer Society Digital Library. IEEE Computer Society.

[8] D. Nguyen, D. Rocker: Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics, 18, 12, 1625-1632, 2002.

[9] J. Gui, H. Li: Penalized Cix regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 21, 3001-3008, 2005.

[10] I. Guyon, J. Weston, S. Barnhill: Gene selection for cancer classification using support vector machines. Machine Learning, 46, 1-3, 389-422, 2002

[11] Y. Ding, D. Wilkins: Improving the performance of SVM-RFE to select genes in microarray data. BMC Bioinformatics, 7(Suppl 2), 1-8, 2006.

[12] S. Shevade, S. Keerthi: A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, 19, 17, 2246-2253, 2003.

[13] G. Cawley, N. Talbot: Gene selection in cancer classification using sparse logistic regression with bayesian regularization. Bioinformatics, 22, 2348-2355, 2006.

[14] L. Ein-Dor, I. Kela, G. Getz, D. Givol, E. Domany: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, 21, 171-178, 2005.

[15] A. kalousis, J. Prados, M. Hilario: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and information systems, 12, 95-116, 2007.

[16] G. Unger, B. Chor: Linear separability of gene expression datasets. IEEE Trans. on computational biology and bioinformatics, Aug., 2008.

[17] R. Tibshirani: Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B: Statist. Methodol. 58, 267-288, 1996.

[18] H. Zou, T. Hastie: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B: Statist. Methodol. 67, 301-320, 2005.

[19] H. Zou: The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429, 2006.

[20] H. Zou, H. Zhang: On the adaptive elastic-net with a diverging number of parameters. The Annals of statistics, 37, 4, 1733-1751, 2009.

[21] M. Yuan, Y. Lin: Model selection and estimation in regression with grouped variables. JRSSB, 68, 49-67, 2006.

[22] L. Meier, S. Geer, P. Buhlmann: The group lasso for logistic regression. JRSSB, 70, 53-71, 2008.

[23] D. Donoho, J. Jin: Higher criticism thresholding: optimal feature selection when useful features are rare and weak. Proc. Natl. Acad. Sci. USA, 105, 14790-14795, 2008.

[24] J. Jin: Impossibility of successful classification when useful features are rare and weak. Proc. Natl. Acad. Sci. USA, 106, 8859-8864, 2009.

[25] R. De, A. Ghosh: Interval based fuzzy systems for identification of important genes from microarray gene expression data: application to carcinogenic development. Journal of Biomedical Informatics, online available, Jul.2009.

[26] Y. Yang, J. Pedersen: A comparative study on feature selection in text categorization. Proc. of the 14th ICML, 412-420, 1997.

[27] A. Dasgupta, P. Drineas, B. Harb: Feature selection methods for text classification. Proc. of KDD, San Jose, CA, USA, 2007.

[28] T. Jirapech-Umpai, S. Aitken: Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6, 148:1-11, 2005.

[29] T. Mitchell: Machine learning. McCraw Hill, 1996.

[30] T. Golub et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531-537, 1999.

[31] V. Tusher, R. Tibshirani, G. Chu: Significance analysis of microarray applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA, 98, 9, 5116-5121, 2001.

[32] L. Yu, C. Ding, S. Loscalzo: Stable feature selection via dense feature groups. Proc. of SIG KDD, Las Vegas, Nevada, USA, 803-811, 2008.

[33] S. Loscalzo, L. Yu, C. Ding: Consensus group stable feature selection. Proc. of SIG KDD, Paris, France, 567-575, 2009.

[34] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA, 96, 6745-6750, 1999.

[35] M. West et al.: Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA, 98, 20, 11462-11467, 2001.

[36] H. Kishino, P. Waddell: Correspondence analysis of genes and tissue types and finding genetic links from microarray data. Genome information, 11, 83-95, 2000.

[37] E. Feng, M. Ng:On sparse Fisher discriminant method for microarray data analysis. Bioinformation, 2(5), 230-234, 2007.

[38] C. Bolmont, A. Lilienbaum, D. Paulin, J. Grimaud: Expression of desmin gene in skeletal and smooth muscle by in situ hybridization using a human desmin gene probe. Journal of Submicrosc Cytol Pathol., 22(1), 117-122, 1990.

[39] Y. Li, C. Campbell, M. Tipping: Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics, 18, 1332-1339, 2002.

[40] L. Young, S. Sanduja, K. Bemis-Standoli, E. Pena, R. Price, D. Dixon: The mRNA binding protiens HuR and tristetraprolin regulate cyclooxygenase 2 expression during colon carcinogenesis. Gastroenterology, 136(5), 1669-1679, 2009.

[41] U. Knippschild, S. Wolff, G. Giamas, C. Brockschmidt, M. Wittau, P. Wai, T. Eismann, M. Stier: The role of the casein kinase 1 family in different signaling pathways linked to cancer development. Onkologie, 28, 508-514, 2005.

[42] L. Kaufman, P. Rousseeuw: Finding groups in data: an introduction to cluster analysis, Wiley, 1990.

[43] A. Strehl: Relationship-based clustering and cluster ensembles for high-dimensional data mining. Ph.D thesis, The University of Texas at Austin, 2002.

TABLE V
SIGNIFICANT GENES OF LEUKEMIA DATA SELECTED BY THE PROPOSED SGS METHOD

CI	GeneID	Gene Discription
16	Y00787	INTERLEUKIN-8 PRECURSOR
	M80254	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR
	L08246	INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1
	M28130	Interleukin 8 (IL8) gene
	M69043	MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3
19	L47738	Inducible protein mRNA
	U35451	Heterochromatin protein p25 mRNA
	M13792	ADA Adenosine deaminase
	M12886	TCRB T-cell receptor, beta cluster
	M14483	PTMA gene extracted from Human prothymosin alpha mRNA
	X00437	TCRB T-cell receptor, beta cluster
21	U05259	MB-1 gene
	D26156	Transcriptional activator hSNF2b
	U29175	Transcriptional activator hSNF2b
	Y08612	RABAPTIN-5 protein
	M27830	AFFX-M27830-5-at (endogenous control)
	M11722	Terminal transferase mRNA
26	U22376	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds
	Z15115	TOP2B Topoisomerase (DNA) II beta (180kD)
	X15949	IRF2 Interferon regulatory factor 2
	Z69881	Adenosine triphosphatase, calcium
32	U50136	Leukotriene C4 synthase (LTC4S) gene
	M84526	DF D component of complement (adipsin)
	X17042	PRG1 Proteoglycan 1, secretory granule
	U46751	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
33	M55150	FAH Fumarylacetoacetate
	X95735	Zyxin
	M16038	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
	U82759	GB DEF = Homeodomain protein HoxA9 mRNA
	M23197	CD33 CD33 antigen (differentiation antigen)
	M27891	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
	M62762	ATP6C Vacuolar H+ ATPase proton channel subunit
	M81695	ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)
	X04085	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)
35	X59417	PROTEASOME IOTA CHAIN
	S50223	HKR-T1
	M31303	Oncoprotein 18 (Op18) gene
38	M63138	DDC Dopa decarboxylase (aromatic L-amino acid decarboxylase)
	M57710	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
	M19045	LYZ Lysozyme
	M83652	PFC Properdin P factor, complement
42	M31211	MYL1 Myosin light chain (alkali)
	X74262	RETINOBLASTOMA BINDING PROTEIN P48
	U32944	Cytoplasmic dynein light chain 1 (hdlc1) mRNA
	X63469	GTF2E2 General transcription factor TFIIE beta subunit, 34 kD
	M91432	ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain
	U20998	SRP9 Signal recognition particle 9 kD protein
	U26266	DHPS Deoxyhypusine synthase
	M2969	IL7R Interleukin 7 receptor

[44] T. Attwood, D. Smith: Introduction to bioinformatics. Prentice Hall, 1999.
[45] L. Jacob, G. Obozinski, J. Vert: Group lasso with overlap and graph lasso. In Proc. of the 26th ICML, Montreal, Canada, 2009.
[46] T. Park, G. Casella: The Bayesian Lasso. Journal of the American Statistical Association, 103, 482, 681-686, 2008.
[47] B. Scholkopf, C. Burges, A. Smola: Advances in kernel methods: support vector learning. MIT Press, Cambridge, MA, 1999.
[48] G. Shakhnarovich, T. Darrell, P. Indyk: Nearest-Neighbor methods in learning and vision. The MIT Press, 2005.
[49] D. Hosmer, S. Lemeshow: Applied logistic Regression, 2nd ed.. New York; Chichester, Wiley, 2000.