Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification

Article *in* IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society · December 2008

DOI: 10.1109/TITB.2008.920787 · Source: PubMed

Yuchun Tang

60 PUBLICATIONS **1,278** CITATIONS

SEE PROFILE

Yichuan Zhao

Georgia State University

91 PUBLICATIONS **497** CITATIONS

SEE PROFILE

Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification

Yuchun Tang, *Member, IEEE*, Yan-Qing Zhang, *Member, IEEE*, Zhen Huang, Xiaohua Hu, *Member, IEEE*, and Yichuan Zhao

***Abstract*—A typical microarray gene expression dataset is usually both extremely sparse and imbalanced. To select multiple highly informative gene subsets for cancer classification and diagnosis, a new *Fuzzy Granular Support Vector Machine—Recursive Feature Elimination* algorithm (FGSVM-RFE) is designed in this paper. As a hybrid algorithm of statistical learning, fuzzy clustering, and granular computing, the FGSVM-RFE separately eliminates irrelevant, redundant, or noisy genes in different granules at different stages and selects highly informative genes with potentially different biological functions in balance. Empirical studies on three public datasets demonstrate that the FGSVM-RFE outperforms state-of-the-art approaches. Moreover, the FGSVM-RFE can extract multiple gene subsets on each of which a classifier can be modeled with 100% accuracy. Specifically, the independent testing accuracy for the prostate cancer dataset is significantly improved. The previous best result is 86% with 16 genes and our best result is 100% with only eight genes. The identified genes are annotated by Onto-Express to be biologically meaningful.**

***Index Terms*—Cancer classification, fuzzy C-means clustering, gene selection, granular computing, microarray gene expression data analysis, recursive feature elimination (RFE), relevance index (RI), support vector machines (SVMs).**

I. Introduction

GENE expression microarrays (including the cDNA microarray and the GeneChip) have been developed as a powerful technology for functional genetics studies that simultaneously measures the mRNA expression levels of thousands to tens of thousands of genes. A typical microarray expression experiment monitors the expression level of each gene multiple times under different conditions or in different phenotypes. For example, a comparison can be made between a healthy tissue and a cancerous tissue, or a tissue of one kind of cancer versus that of another. By collecting such huge gene expression datasets, it opens the possibility to distinguish phenotypes and to identify disease-related genes whose expression patterns are excellent diagnostic indicators [1].

One of the goals of microarray data analysis is cancer classification. For example, one well-known challenge using gene expression microarray data is to distinguish between two variants of leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [2]. The AML/ALL challenge can be modeled as a binary classification problem. If a sample is ALL, it is classified as negative or -1, otherwise it is positive or 1. This strategy can be applied to many other microarray datasets.

A typical microarray gene expression dataset is extremely "sparse" in the sense that the dataset usually comes with only dozens of tissue samples but with thousands or even tens of thousands of gene features. In a high-dimensional space with each gene feature as one dimension, we can only find a few of points that represent the tissue samples. This extreme sparseness is believed to significantly deteriorate the performance of cancer classification. As a result, the ability to extract informative genes while removing irrelevant or redundant genes is crucial for accurate classification. Furthermore, it is also helpful for biologists to find the inherent cancer-causing mechanism, and hence to develop better diagnostic methods or find better therapeutic treatments [2], [3].

As a brief summary, there are two related challenging tasks for data mining and bioinformatics scientists.

1) Gene Subset Extraction: Given some tissues, extract cancer-related genes while removing irrelevant or redundant genes. Biologically, genes are expected to regulate cancers in complex and nonindependent ways. As a result, it is more desirable to extract cancer-related genes together as a group than to extract them one by one.
2) Cancer Classification: Given a new tissue, predict if it is healthy or not; or categorize it into a correct class.

The rest of the paper is organized as follows. In Section II, a brief review is given for current computational methods in this field. After that, the new FGSVM-RFE algorithm is presented in Section III. Sections IV and V evaluate the performance of FGSVM-RFE on three public microarray gene expression datasets. Finally, Section VI discusses and concludes the paper.

II. Background and Related Work

A. SVM for Cancer Classification

Based on [3], a support vector machine (SVM) is adopted for cancer classification in this paper. SVM is a new-generation

Manuscript received November 1, 2007; revised January 11, 2008; accepted February 15, 2008. First published March 21, 2008; current version published November 5, 2008. The work of Z. Huang was supported in part by the National Institutes of Health under Grant NIAID AI058051 and in part by the Southeastern Center for Emerging Biologic Threats (Emory University). The work of X. Hu was supported in part by the National Science Foundation under Career Grant IIS 0448023 and Grant CCF 0514679, and in part by the Pennsylvania Department of Health Grants.

Y. Tang was with the Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994 USA. He is now with the Secure Computing Corporation (formerly CipherTrust Inc.), Alpharetta, GA 30022 USA (e-mail: tyczjs@yahoo.com).

Y.-Q. Zhang is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994 USA (e-mail: yzhang@cs.gsu.edu).

Z. Huang is with the Department of Chemistry and Department of Biology, Georgia State University, Atlanta, GA 30303-3083 USA (e-mail: huang@gsu.edu).

X. Hu is with the College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA, and also with the School of Computer & Information Engineering, Henan University, Henan, China (e-mail: Tony.Hu@ischool.drexel.edu).

Y. Zhao is with the Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303-3083 USA (e-mail: matyiz@langate.gsu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2008.920787

learning system based on recent advances in statistical learning theory [4]. Due to extreme sparseness of microarray gene expression data, the dimension of input space is already high enough so that the cancer classification is already as simple as a linear separable task [3]. It is unnecessary and even useless to transfer it to a higher implicit feature space with a nonlinear kernel. As a result, in this paper, we adopt the linear kernel as shown in (1) as the cancer classifier

$$K(x_1, x_2) = \langle x_1, x_2 \rangle \tag{1}$$

where x_1 and x_2 are points in a d-dimensional Euclidian space, and $\langle x_1, x_2 \rangle$ denotes the inner product of x_1 and x_2.

For a linear kernel SVM, the margin width is $2/\|w\|$, where w can be calculated by

$$w = \sum_{i=1}^{N_s} \alpha_i y_i x_i \tag{2}$$

where N_s is the number of support vectors, which are the training samples with $0 < \alpha_i \leq C$. Note that C is a "regulation parameter" used to tradeoff the training accuracy and model complexity so that a superior generalization capability can be achieved. SVM is believed to be a superior model for sparse classification problems compared to other models. However, the sparseness of microarray data is so extreme that even an SVM classifier is unable to achieve a satisfactory performance. A preprocessing step for gene selection can help for a more reliable classification. Currently, there are mainly two kinds of gene selection algorithms: correlation-based gene ranking algorithms and backward elimination algorithms.

B. S2N for Gene Selection

Correlation-based feature ranking algorithms work in a forward selection way by ranking genes individually in terms of a correlation-based metric, and then, the top-ranked genes are selected to form the most informative gene subset. Signal-to-noise (S2N) [5], as shown in (3), is one of the most popular correlation-based ranking metrics

$$w_i = |\mu_i(+) - \mu_i(-)|/(\delta_i(+) + \delta_i(-)) \tag{3}$$

where $\mu_i(+)$ and $\mu_i(-)$ are the mean values of the ith gene's expression data over positive and negative samples in the training dataset, respectively. $\delta_i(+)$ and $\delta_i(-)$ are the corresponding standard deviations. A larger w_i means that the ith gene is more informative for cancer classification. The S2N algorithm, as well as other similar algorithms, is straightforward to understand and works efficiently. However, a common drawback is that these algorithms implicitly assume that genes are orthogonal to each other, and thus, can only detect relations between class labels and a single gene. The mutual information such as redundancy or complementariness among multiple genes cannot be detected.

C. SVM-RFE for Gene Selection

On the other hand, backward elimination algorithms iteratively remove one "least important" gene at one time. Then, the remaining genes are ranked again. Recently, backward elimination algorithms have achieved notable performance improvement over the previous methods [3], [6]. The reason of the good performance of backward elimination algorithms is that they do not make the orthogonality assumption, and thus, can consider multiple gene features simultaneously. However, backward elimination algorithms rank all remaining genes in one list. In this paper, we find that a microarray gene expression dataset is usually (highly) imbalanced between genes with potentially different biological functions. If genes for one specific function are scarce, all of them may be low-ranked in the list, and hence eliminated. This may induce information loss.

One well-known backward elimination algorithm is the support vector machine—recursive feature elimination algorithm (SVM-RFE) [3]. In SVM-RFE, the removed gene should change the objective function J in (4) least

$$J = \|w\|^2/2 \tag{4}$$

where w is calculated by (2), because a linear SVM is used.

The optimal brain damage (OBD) algorithm [7] approximates the change of J by removing the ith gene by expanding J in Taylor series to the second-order

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2. \tag{5}$$

At the optimum of J, the first order is neglected and the second order becomes

$$\Delta J(i) = (\Delta w_i)^2. \tag{6}$$

Because removing the ith gene means $\Delta w_i = w_i$, we can adopt w_i^2 as the ranking criteria. The gene with the smallest w_i^2 has the smallest effect on classification, and hence is removed. In practice, more than one gene could be removed in one step. A parameter f decides how many genes are removed in one step. If $0 < f < 1$, a fraction of f bottom-ranked genes is removed at each step. If $f = -1$, only one gene is removed. In each step, a new linear SVM is trained in a smaller feature space, and thus, each remaining gene is assigned a new weight w_i^2 to be ranked again. This process is repeated until the predefined number of features remain. The SVM-RFE with $f = -1$ is the most time-consuming because maximum steps are needed in this case.

III. FGSVM-RFE Algorithm for Gene Selection

The FGSVM-RFE is based on principles of granular computing [8]–[10]. It works in three stages. At the first stage, two relevance index (RI) metrics are proposed to eliminate most irrelevant genes and to "rebalance" distribution between positive-related genes and negative-related genes. At the second stage, the remaining genes are recursively grouped into different function granules by the fuzzy C-means clustering algorithm (FCM) [11]. As a result, redundant genes can be eliminated based on SVM ranking in each function granule. Furthermore, informative genes in different function granules can be extracted in balance. Finally, one more SVM-based ranking is executed to finely select the final informative gene subset.

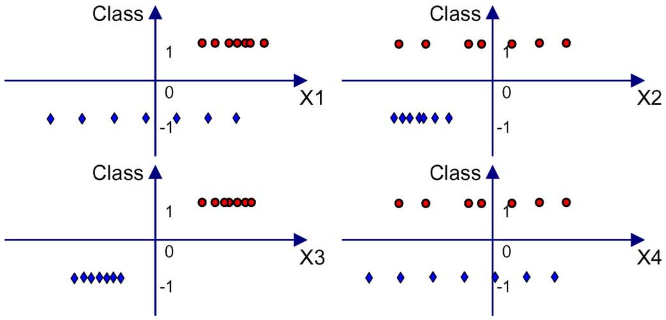

Fig. 1. Positive-related gene, negative-related gene, both, neither.

A. Relevance Index

The "RI" was used to measure the relevance of a feature to a cluster in [12] to ease an unsupervised clustering process. A similar idea is used as the first step of the FGSVM-RFE. Because a gene may be negative-related or positive-related, (7) and (8) define the negative RI and the positive RI to measure the negative correlation and the positive correlation of a gene with the cancer being studied, respectively. The negative and positive RI metrics are used to prefilter most of the irrelevant genes and to rebalance gene distribution to ease the following gene selection and supervised classification

$$R_{i-} = 1 - \lambda_{i-}^2/\lambda_i^2 \tag{7}$$

$$R_{i+} = 1 - \lambda_{i+}^2/\lambda_i^2 \tag{8}$$

where λ_i^2, λ_{i-}^2, and λ_{i+}^2 are the variances of the projected values on the ith gene of the whole training samples, the negative training samples, and the positive training samples, respectively.

For example, gene X1 in Fig. 1 is positive-related because the local variance among positive samples is much smaller than the global variance on the whole samples. Notice that a circle denotes a positive sample and a diamond denotes a negative sample. Similarly, gene X2 is negative-related, gene X3 is both negative-related and positive-related, and gene X4 can be viewed as an "irrelevant" gene in that it is neither negative-related nor positive-related. Notice also that variance and the direction of regulation are independent. For example, gene X2 would still be negative-related if the diamonds were on the right side (higher expression than the mean).

To apply RI metrics for gene selection, a negative filtering threshold $\alpha_- \in [0, 1)$ and a positive filtering threshold $\alpha_+ \in [0, 1)$ need to be decided. The ith gene is "negative-related" if $R_{i-} \geq \alpha_-$ because the projections of the negative samples are closer than the projections of the whole samples on the ith gene. Similarly, it is "positive-related" if $R_{i+} \geq \alpha_+$. If $R_{i-} < \alpha_-$ and $R_{i+} < \alpha_+$, it is "irrelevant." A gene may be both negative-related and positive-related. These two filtering thresholds should be selected carefully. First, the thresholds can not be too large; otherwise, information loss may happen because some cancer-related genes may be eliminated by mistake. Second, negative-related genes and positive-related genes should be selected in balance; otherwise, the minor genes may be totally eliminated to result in performance degradation, especially when negative-related genes and positive-related genes are significantly imbalanced in the original dataset.

B. Fuzzy C-Means Clustering

RI alone cannot extract good gene subsets. The shortcoming of RI is that it assumes the independence between different genes. As we know, the assumption is not true for microarray gene expression data. If the filtering thresholds are too large, many informative genes may be unfortunately eliminated.

Genes functioning similarly might have similar patterns of expression. As a result, if genes with similar expression patterns are grouped together into clusters, a few typical genes in a cluster may be selected and other genes in the cluster may be safely eliminated without significant information loss. On the other hand, an informative gene may contribute to cancer classification with multiple biological functions. By being clustered into multiple function granules, the informative gene may be ranked low in some granules but ranked high in other granules, and hence gains more opportunities to be extracted. Therefore, after the prefiltering by RI metrics, the FCM is adopted to group genes into different function clusters. Another potential advantage with this fuzzy clustering is that genes regulating cancers in different ways can be extracted in balance.

For gene selection, first the expression data matrix is transposed so that each gene is a row and each tissue is a column. And then, the FCM groups genes into K clusters with centers $c_1, \ldots, c_k, \ldots, c_K$ in the training tissue samples space. FCM assigns a real-valued vector $U_i = \{\mu_{1i}, \ldots, \mu_{ki}, \ldots, \mu_{Ki}\}$ to each gene. $\mu_{ki} \in [0, 1]$ is the membership value of the ith gene in the kth cluster. Initially, the centers are randomly decided and every $\mu_{ki} = 1/k$. The larger membership value indicates the stronger association of the gene to the cluster. Membership vector values μ_{ki} and cluster centers c_k can be obtained by minimizing

$$J(K, m) = \sum_{k=1}^{K} \sum_{i=1}^{N} (\mu_{ki})^m \, d^2(x_i, c_k) \tag{9}$$

$$d^2(x_i, c_k) = (x_i - c_k)^T A_k (x_i - c_k) \tag{10}$$

$$\sum_{k=1}^{K} \mu_{ki} = 1, 0 < \sum_{i=1}^{N} \mu_{ki} < N. \tag{11}$$

In (9), K and N are the number of clusters and the number of genes, respectively. $m > 1$ is used to control the "fuzziness" of the resulting clusters, μ_{ki} is the degree of membership of the ith gene in the kth cluster, and $d^2(x_i, c_k)$ is the square of distance from the ith gene to the center of the kth cluster. In (10), $1 \leq k \leq K$ and $1 \leq i \leq N$, and A_k is a symmetric and positive definite matrix. In this paper, we use the identity matrix as A_k so that $d^2(x_i, c_k)$ corresponds to the square of the Euclidian distance. Equation (11) indicates that empty clusters are not allowed. Please refer to [11] for more details about the FCM. An informative gene may not be "crisply" negative-related/positive-related, but may regulate two classes to different extents. And hence, a finer granularity than binary split is needed. On the other hand, it is expensive for gene ranking if too many clusters are generated. In the second stage of the FGSVM-RFE, genes are roughly grouped into five granules. In each step, clusters are recombined after gene reduction, and then, the genes are reclustered. Our empirical studies demonstrate that highly informative gene subsets can be extracted with this process.

C. FGSVM-RFE for Gene Selection

One of our contributions to this gene selection task is to propose categorizing genes into the following four groups and deal with each group separately:

1) informative genes, which are essential for cancer classification and cancer diagnosis;
2) redundant genes, which may also be helpful for cancer classification but some informative genes regulate cancers similarly but more significantly;
3) irrelevant genes, which are not cancer-related
4) noisy genes, which are not cancer-related but they have negative effects on cancer classification.

Usually, for gene selection, the first step is to rank genes, and then, a small subset of top-ranked genes is selected. Therefore, an ideal algorithm should rank the first group of genes close to the top while ranking the last three groups of genes close to the bottom. However, it is difficult to achieve an ideal ranking. First, the inherent cancer-related factors are very likely mixed with other noncancer-related but biological-experiment-related factors for classification. Second, some biological-experiment-related factors may even have more significant effects on classifying the training dataset. It is actually the notorious "overfitting" problem. It is even worse when the training dataset is too small to embody the inherent cancer-related data distribution, which is common for microarray gene expression data analysis. We believe that noisy genes of the fourth group play the key role to hide the inherent cancer-related distribution and to confuse a classifier. Briefly, to say, a noisy gene may have three kinds of negative effects on cancer classification.

1) A noisy gene may discriminate the training samples by some biological-experiment-related factors so that the noise gene is ranked high.
2) A noisy gene or a group of noisy genes may be complementary to some redundant or irrelevant genes so that these redundant or irrelevant genes are ranked high.
3) A noisy gene or a group of noisy genes may conflict with some informative genes so that these informative genes are ranked low.

As a result, the inherent cancer-related distribution is blurred by noise to induce information loss in the process of gene selection. Furthermore, imbalanced gene distribution may also induce information loss as analyzed earlier.

The new FGSVM-RFE algorithm is designed to extract highly informative gene subsets by decreasing information loss. It works in three stages. Fig. 2 shows a sketch of FGSVM-RFE. At the first stage, genes are coarsely grouped into "relevant granule" and "irrelevant granule" based on RI metrics. The relevant granule consists of negative-related genes and positive-related genes, while the irrelevant granule is comprised of irrelevant genes with both small R_{i+} value and small R_{i-} value. Only genes in the relevant granule survive for the following stages. As a result, we can avoid highly ranking irrelevant genes due to undesirable correlations between them and noisy genes. This prefiltering process can dramatically decrease the number of candidate genes on which FCM works. Therefore, it can improve both the effectiveness and the efficiency of the following stages of FGSVM-RFE.

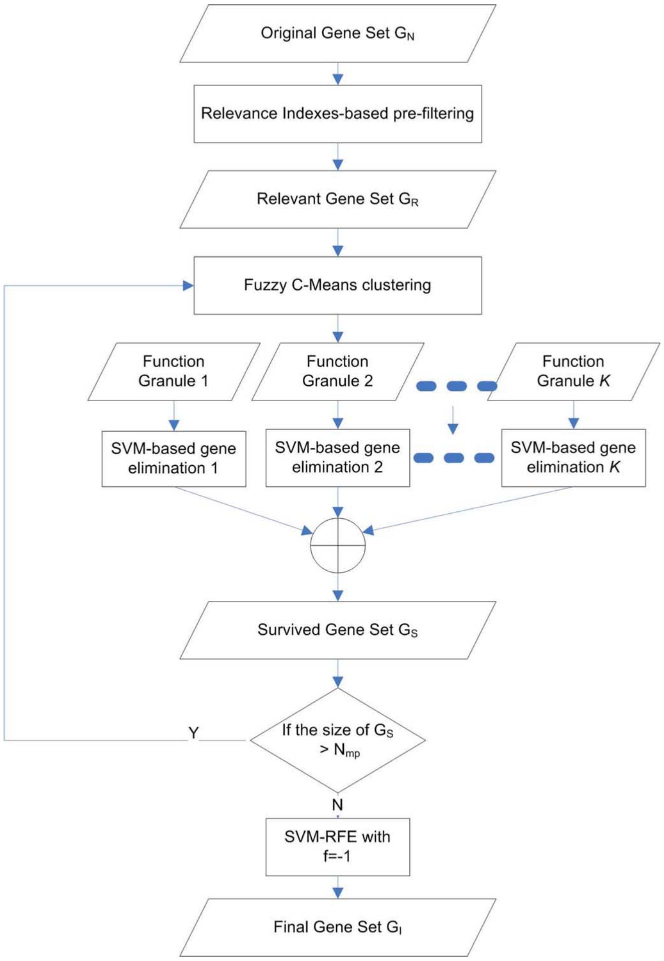

Fig. 2. FGSVM-RFE.

At the second stage, in each step, the survived genes are clustered by the FCM into five function granules based on their expression patterns. In each function granule, a linear SVM is modeled to rank genes, as shown in (4)–(6). The lower-ranked genes are removed as redundant genes and the higher-ranked genes are selected as new survived genes. And then, all survived genes in these function granules are disjunctively combined into the next step. This process is repeated until the number of survived genes is less than or equal to a prespecified threshold N_{mp}. As a result, we can decrease noise between redundant genes and noisy genes. Furthermore, due to complex correlations among genes, the similarity is by no means a "crisp" concept. The FCM deals with that by assigning a gene into several clusters with different membership values. Therefore, a really informative gene is more likely to survive.

TABLE I
CHARACTERISTICS OF DATASETS USED IN EXPERIMENTS

dataset			#genes	#samples	#neg : #pos
Prostate cancer	training	[13]	12600	102	52:50
	testing	[14]		34	25:9
AML/ALL	training	[2]	7129	38	27:11
	testing	[2]		34	20:14
Colon cancer		[15]	2000	62	40:22

At the third stage, because most of irrelevant and redundant genes have been removed, one more SVM-based ranking is executed to finely select the final "most informative" gene subsets. Similar to [3], SVMs modeled on these final gene subsets are used for cancer classification.

IV. EXPERIMENT DESIGN

A. Datasets and Experimental Environment

Table I lists characteristics of three datasets used in the experiments. All datasets are available at http://sdmc.lit.org.sg/GEDatasets/Datasets.html. The original data sources are also referred to in Table I. The experiments are carried out on a machine with a Pentium M CPU at 1.73 GHz and 1 GB of memory. The software is based on the OSU SVM Classifier Matlab Toolbox (available at http://www.ece.osu.edu/ maj/osu_svm/), which implements a Matlab interface to LIBSVM (available at http://www.csie.ntu.edu.tw/cjlin/libsvm).

B. Data Modeling

In a similar way to the treatment by Golub *et al.* [2], the original dataset is simply normalized by decreasing the mean of the corresponding gene vector from each gene's expression data and then dividing by the corresponding standard deviation. As a result, each gene vector has a mean of 0 and a standard deviation of 1. To avoid overfitting, the mean and standard deviation are calculated by using the training dataset. If leave-one-out cross-validation (LOOCV) is used, the validation samples are kept out from calculating these two values.

LOOCV leaves out a single sample of the data, builds a classifier on the remaining samples, and then, evaluates classification performance on the test sample. This step is repeated for each sample to obtain the aggregated classification performance.

As the stopping criterion of the FCM, the maximal iteration number is 100, and the minimal improvement $\epsilon = 10^{-5}$. The "fuzziness degree" $m = 1.15$ so that each gene has strong association with two clusters [13]. In the second stage of the FGSVM-RFE, at each step, the survived genes are grouped into five function granules, in each of which a linear SVM with $C = 1$ is modeled for gene ranking and 50% lower-ranked genes are removed ($f = 0.5$), and then, all of the five subsets of survived genes are disjunctively combined for reclustering in the next step.

Notice in each step, the fuzzy membership values are defuzzified in such a way that a gene is always assigned into two granules with the top two largest membership values. The assumption is that different gene function groups are clustered based on their expression strengths. Some genes whose expression strengths are between two groups may be better to be clustered into the two groups at the same time. In this way, each gene has two opportunities to be extracted.

This process is repeated until the number of survived genes is less than or equal to N_{mp}. The value of N_{mp} depends on the practical utilities of the extracted gene subset. A small gene subset is desirable for further biological study; on the other hand, it may result in information loss if too many genes are discarded in the second stage. We select $N_{mp} = 64$ here. At the third stage, a linear SVM is modeled for ranking all survived genes and only one gene is removed in each step ($f = -1$). As a result, nested gene subsets from (at most) 64 genes to 1 gene are extracted.

Due to page limit, we only report the experimental results on the prostate cancer dataset. Please refer http://tinman.cs.gsu.edu/cscyntx/fgsvm_full.pdf for results on two other datasets.

V. PROSTATE CANCER DATA

A. Data Description

The prostate cancer dataset is used for tumor versus normal classification. The training dataset, from http://www-genome.wi.mit.edu/mpr/prostate, consists of 102 prostate samples (52 with tumors and 50 without tumors); while the testing dataset, from http://carrier.gnf.org/welsh/prostate/, has 34 samples (25 with tumors and 9 without tumors). These two datasets are prepared under different biological experimental contexts. There is a nearly 10-fold difference in overall microarray intensity between them (not i.i.d.). The 12 600 features correspond to some normalized gene expression values extracted from the microarray image.

Here, negatives are defined to be tumor samples, while positives are normal prostate samples without tumor. The genes distribution in the prostate cancer dataset is highly imbalanced between negative-related genes and positive-related genes. If $\alpha_- = \alpha_+ = 0.5$, 4761 negative-related genes and only 110 positive-related genes are selected. To alleviate the imbalance, $\alpha_- = 0.75$ and $\alpha_+ = 0.5$ are used to select 721 negative-related genes and 110 positive-related genes. There is no overlapping between positive-related genes and negative-related genes.

B. Result Analysis

We run FGSVM-RFE 20 times. For each run, the "best" performance on nested gene subsets from (at most) 64 genes to 1 gene is reported. Under different validation heuristics, the best performance may be observed at different numbers of genes. As it is more reliable as explained earlier, the gene subset with the highest LOOCV accuracy on the original training dataset is reported. If there is a tie, the gene subset with the highest prediction accuracy on the original testing dataset is reported. Averaged on 20 runs, the FGSVM-RFE can find a gene subset with 99.71% testing accuracy with at most ten genes.

The FGSVM-RFE extracts three gene subsets with 100% LOOCV accuracy on the original training dataset and also 100% prediction accuracy on the testing dataset, which are reported in

TABLE II
PERFORMANCE COMPARISON ON THE PROSTATE CANCER DATASET

models		leave-1-out validation on the original training dataset	prediction on the original testing dataset
Bagging C4.5	[17]	N/A	73.53% (3071)
S2N+k-NN	[13]	90% (4)	86% (16)
GS1+k-NN	[18]	95.1% (8)	N/A
FGSVM-RFE		100% (8) 100% (17) 100% (15,14)	100% (8) 100% (17) 100% (15,14)

TABLE III
FIRST HIGHLY INFORMATIVE GENE SUBSET SELECTED BY THE FGSVM-RFE ON THE PROSTATE CANCER DATASET

rank/index		Probe ID	Description of Gene Function
1	6185	37639_at^{M}	hepsin
2	4649	32035_at^{BM}	major histocompatibility complex
3	5821	36555_at	synuclein, gamma
4	5045	33744_at^{BC}	gem associated protein 4
5	10537	33121_g_at^{B}	regulator of G-protein signalling 10
6	6368	38300_at^{BM}	frizzled homolog 1
7	11818	914_g_at^{B}	v-ets virus E26 oncogene
8	5402	35178_at^{B}	WNT inhibitory factor 1

TABLE IV
PERFORMANCE COMPARISON ON THE PROSTATE CANCER DATASET

model	#gene	acc(%)	tnr(%)	tpr(%)	time(s)
No selection	12600	32.4	8.0	100	1011
S2N	22	79.4	100	22.2	76
SVM-RFE	28	94.1	92.0	100	69
RSVM-RFE	16	79.4	72.0	100	64
FSVM-RFE	38	76.5	68.0	100	1794
FGSVM-RFE	8	100	100	100	140

Table II. In each cell, the accuracy is followed by the number of genes.

With LOOCV on the training dataset and then prediction on the testing dataset, we observe obvious performance improvements. Tan *et al.* reported 73.53% testing accuracy with 3071 genes [14]. Singh *et al.* reported 90% validation accuracy with four or more genes and 86% testing accuracy with 16 genes [15]. Yang *et al.* reported a 95.1% validation accuracy with eight genes [16]. The FGSVM-RFE generates higher classification accuracy than all of them.

Table III reports the highly informative gene subset with eight genes, on which a linear SVM can be modeled with 100% LOOCV accuracy on the training dataset and also 100% prediction accuracy on the testing dataset. Please refer http://tinman.cs.gsu.edu/cscyntx/fgsvm_full.pdf for the other two gene subsets.

As Onto-Express (http://vortex.cs.wayne.edu) is a novel tool to automatically translate a list of differentially regulated genes into functional profiles and to characterize the condition impact [17], it is used to analyze the identified gene subsets. A set of 31 genes, which is formed by disjunctively combining the three highly informative gene subsets, is submitted to Onto-Express using the initial pool of 12 600 genes as the reference set. As suggested by [17], the hypergeometric distribution is adopted to calculate *p*-value for each gene ontology (GO) term and the false discovery rate (FDR) is used for *p*-value correction. We concentrate on all three primary GO categories including biological processes, molecular functions, and cellular components that are significant at 5% ($p < 0.05$). Seven of eight genes in Table III, or 22 of 31 genes in all, are annotated by Onto-Express with significant biological functions. In the second column of Table III, a gene is superscripted by a ^{B} if it is significantly related to some biological process. Similarly, ^{C} means cellular component, and ^{M} means molecular function. It demonstrates that many identified genes can quickly be assigned biological meanings with Onto-Express.

In a gene regulation network, different gene subsets may regulate cancer in different ways. As a result, the three 100% accurate gene subsets are expected to provide better understanding for the cancer by decreasing information loss. Moreover, genes that survive in multiple subsets deserve higher priority for further cancer study. Although some genes have been already reported to be cancer-related, it is the first time they are reported to coregulate prostate cancer in the same subset. As such, it may be more helpful to detect a coregulation network. For example, genes in the same subset may regulate prostate cancer from the same pathway or with similar/complementary functions.

VI. DISCUSSION AND CONCLUSION

A. *Discussion*

As mentioned earlier, the FGSVM-RFE embeds three assumptions into the gene selection process. First, the RI metrics are assumed to be useful to remove irrelevant genes and corresponding noises. Second, recursive FCM is assumed to be useful to remove redundant genes and corresponding noises. Third, both RI metrics and recursive FCM are useful to select genes from different function groups in balance. The superior performance of the FGSVM-RFE on the three datasets verifies these assumptions.

To show the contribution of each component in the FGSVM-RFE, we conduct another group of experiments. As a typical correlation-based metric, signal to noise (S2N) [5] is used as a "finest granulation" method in the sense that each gene forms a small granule independently. Similarly, the SVM-RFE [3] is a "coarsest granulation" method because all genes form only one granule and implicitly assume that all genes regulate cancers together with some correlations. We also remove recursive FCM and RI metrics from the FGSVM-RFE as two new methods denoted by RSVM-RFE and FSVM-RFE, respectively. These four methods, along with SVM classification without gene selection, are compared to the FGSVM-RFE. The LOOCV on the training dataset is used for modeling and the corresponding testing performance is reported. Three metrics, accuracy (acc), true-negative rate (tnr), true-positive rate (tpr), are reported for each algorithm.

As shown in Table IV, the SVM classification without gene selection is not accurate. For an extremely sparse microarray dataset, it is very likely that a lot of gene features are irrelevant, redundant, or even noisy. With the "finest granulation" S2N method or the "coarsest granulation" SVM-RFE method for gene selection, the classification performance is improved. It proves that both of them are useful to select informative genes from noninformative genes.

Under the principle of granular computing, a better granulation is likely to stay between these two granulation extremes to further decrease information loss while removing noises. Guided by the assumptions, the FGSVM-RFE is designed and achieves better performance than S2N and SVM-RFE. The experiments also demonstrate that, without removing irrelevant genes or without removing redundant genes, either FSVM-RFE or RSVM-RFE performs worse than FGSVM-RFE.

Furthermore, gene distribution between negative-related and positive-related or among multiple function groups may be not balanced on the original dataset. Our granulation method is also useful to pick up more informative genes by "rebalancing" the distribution in the process of gene selection. With recursive fuzzy granulation, minor genes may form a "pure" granule so that at least some of them can be extracted. The balance ability is demonstrated to be critical for the superior performance of FGSVM-RFE. As shown in Table IV, every method except the FGSVM-RFE is imbalanced in that it is prone to the positive (higher tpr) or the negative (higher tnr).

We also compare the speed of these methods. The reported running time includes the time to extract nested gene subsets from 64 genes to 1 gene and LOOCV on these subsets. For SVM classification without gene selection, the running time is taken by the LOOCV on the original gene set. Although slower than S2N and SVM-RFE, the FGSVM-RFE still finishes in a reasonable time. The FGSVM-RFE is even faster than directly running SVM classification on the original gene set. The fast running time of FGSVM-RFE is mainly due to removing a large amount of irrelevant genes with RI metrics. As a proof, the FSVM-RFE runs much slower without RI prefiltering.

Recently, Yousef *et al.*, proposed SVM-RCE [18], which ranks each gene in each individual cluster by the SVM coefficient weights rather than ranking each cluster as a unit. The size of the clusters, rather than the number of clusters, is reduced. They also did not use RI filtering.

B. Conclusion

To extract multiple informative gene subsets for reliable cancer classification, the FGSVM-RFE is designed in this paper. First, the FGSVM-RFE utilizes RI metrics for gene prefiltering to remove most of the irrelevant genes. Second, the FGSVM-RFE explicitly groups genes with similar expression patterns into function granules by recursively clustering with the FCM. An SVM-based ranking is thereafter carried out in each granule to safely remove lower-ranked genes as redundant genes because higher informative genes with similar functions still survive. Finally, the FGSVM-RFE deals with complex correlations among genes by assigning a gene into several granules with different fuzzy membership values so that a really informative gene can achieve more than one opportunity to be extracted.

The FGSVM-RFE can find multiple compact "highly informative" gene subsets on each of which an SVM with 100% accuracy can be modeled. This suggests that different gene subsets may coregulate cancer. All of these highly informative gene subsets are associated with many significant biological functions by Onto-Express. Of course, the newly identified genes by this algorithm need to be further confirmed biologically, which may generate more insights for cancer mechanism, treatment, and study. Nevertheless, the extraction of these cancer-related gene subsets may help to stimulate and guide detailed cancer studies on the gene functions.

REFERENCES

[1] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: Facing the challenges," *SIGKDD Explor.*, vol. 5, no. 2, pp. 1–5, 2003.

[2] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.

[4] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[5] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data.," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[6] F. Li and Y. Yang, "Analysis of recursive gene selection approaches from microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3741–3747, 2005.

[7] Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage," in *Proc. Adv. Neural Inf. Process. Syst. II*, 1990, pp. 598–605.

[8] T. Y. Lin, "Data mining and machine oriented modeling: A granular computing approach," *Appl. Intell.*, vol. 13, no. 2, pp. 113–124, 2000.

[9] W. Pedrycz, *Granular Computing: An Emerging Paradigm*. Heidelberg, Germany: Physica-Verlag, 2001.

[10] Y. C. Tang, B. Jin, and Y.-Q. Zhang, "Granular support vector machines with association rules mining for protein homology prediction," *Artif. Intell. Med.*, vol. 35, no. 1–2, pp. 121–134, 2005.

[11] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition)*. New York: Springer, 1981.

[12] K. Y. Yip, D. W. Cheung, and M. K. Ng, "HARP: A practical projected clustering algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1387–1397, Nov. 2004.

[13] D. Dembélé and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973–980, 2003.

[14] A. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Appl. Bioinf.*, vol. 2, no. 3, pp. S75–S83, 2003.

[15] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[16] K. Yang, Z. Cai, J. Li, and G. Lin, "A stable gene selection in microarray data analysis," *BMC Bioinf.*, vol. 7, no. 228, 2006.

[17] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: Current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[18] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Recursive cluster elimination (rce) for classification and feature selection from gene expression data," *BMC Bioinf.*, vol. 8, no. 144, 2007.

[19] J. B. Welsh, P. P. Zarrinkar, L. M. Sapinoso, S. G. Kern, C. A. Behling, B. J. Monk, D. J. Lockhart, R. A. Burger, and G. M. Hampton, "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," *Cancer Res.*, vol. 61, no. 16, pp. 5974–5978, 2001.

[20] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 96, no. 12, pp. 6745–6750, 1999.

Yuchun Tang (S'04–M'07) received the B.S. degree in computer science from Civil Aviation University of China, Tianjin, China, in 1996, the M.S. degree in optical and electrical engineering from the Beijing Institute of Technology, Beijing, China, in 1999, and the Ph.D. degree in computer science from Georgia State University (GSU), Atlanta, in 2006.

From 1999 to 2001, he was a Research Scientist at the Institute of Computer Science and Technology, Beijing University, Beijing. During the summer of 2003, he was a Research Visitor with the Berkeley Initiative in Soft Computing (BISC) Program, University of California at Berkeley, Berkeley. From 2004 to 2005, he was a Research Fellow with the Molecular Basis for Disease Program, GSU. He is currently a Principal Research Scientist in the Secure Computing Corporation (former CipherTrust Inc.), Alpharetta, GA. His current research interests include knowledge discovery and data mining, machine learning, statistical learning, computational intelligence, soft computing, granular computing, text mining, artificial intelligence, intelligent data analysis, and decision support systems. He has been applying the aforesaid techniques in many domains including bioinformatics, medical informatics, computational biology, computational chemistry, Web information retrieval and information extraction, Internet message security and spam filtering, business intelligence, etc.

Dr. Tang was the recipient of an outstanding research award at GSU in 2004.

Yan-Qing Zhang (S'94–M'97) received the B.S. and M.S. degrees in computer science and engineering from Tianjin University, Tianjin, China, in 1983 and 1986, respectively, and the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, in 1997.

He is currently an Associate Professor in the Department of Computer Science, Georgia State University, Atlanta. His current research interests include hybrid intelligent systems, data mining, bioinformatics, medical informatics, computational Web intelligence, computational intelligence, granular computing, and statistical learning. He is the author or coauthor of more than 50 journal papers and 110 conference papers. He is the author of two books, Co-Editor of three books, and has published 12 book chapters. He has served as a Reviewer for about 50 international journals.

Dr. Zhang is a member of the Bioinformatics and Bioengineering Technical Committee, Granular Computing Technical Committee, and Data Mining Technical Committee of the Computational Intelligence Society of IEEE and the Technical Committee on Pattern Recognition for Bioinformatics of the International Association of Pattern Recognition. He is a Program Co-Chair and Bioinformatics Track Chair of the IEEE 7th International Conference on Bioinformatics & Bioengineering, a Program Co-Chair of the 2006 IEEE International Conference on Granular Computing and the 2005 IEEE-ICDM Workshop on MultiAgent Data Warehousing and MultiAgent Data Mining. He is a Guest Co-Editor for the Special Issue on Computational Intelligence Approaches in Computational Biology and Bioinformatics of the IEEE/ACM Transactions on Computational Biology and Bioinformatics. He is an Associate Editor of the *Journal of Computational Intelligence in Bioinformatics and Systems Biology*, and an Editorial Board member of the *International Journal of Data Mining and Bioinformatics*. He has served as a Committee Member in about 100 international conferences.

Zhen Huang received the B.S. degree in analytical chemistry from Sichuan University, Sichuan, China, the M.S. degree in organic chemistry from Peking University, Beijing, China, and the Ph.D. degree in bio-organic chemistry from the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, in 1984, 1987, and 1994, respectively, all in chemistry.

Since 2004, he has been an Associate Professor of Biological Chemistry in the Department of Chemistry, Georgia State University, Atlanta. His current research interests include selenium derivatization of nucleic acids for X-ray crystallography, and RNA microchip development for gene expression profiling.

Xiaohua Hu (M'01) received the B.Sc. (Software) degree from Wuhan University, Wuhan, China, in 1985, the M.Eng. degree in computer engineering from the Institute of Computing Technology, Chinese Academy of Science, Beijing, in 1988, the M.Sc. in computer science from Simon Fraser University, Burnaby, BC, Canada, in 1992, and the Ph.D. degree in computer science from the University of Regina, Regina, SK, Canada, in 1995.

He is currently an Associate Professor and Founding Director of the Data Mining and Bioinformatics Laboratory, College of Information Science and Technology, Drexel University, Philadelphia, Pennsylvania. His research projects are funded by the National Science Foundation (NSF), the U.S. Department of Education, and the Pennsylvania Department of Health. He is the author or coauthor of more than 150 peer-reviewed research papers in various journals, conferences, and books such as various IEEE/ACM Transactions. He is the Co-Editor of ten books/proceedings. His current research interests include biomedical literature data mining, bioinformatics, text mining, semantic web mining and reasoning, rough set theory and application, and information extraction and information retrieval.

Dr. Hu is the Founding Editor-in-Chief of the *International Journal of Data Mining and Bioinformatics*. He was the recipient of several awards including the 2005 NSF Career Award, the Best Paper Award at the 2007 International Conference on Artificial Intelligence, and the Best Paper Award at the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.

Yichuan Zhao received the B.S. degree in mathematics and the M.S. degree in applied mathematics from Peking University, Beijing, China, in 1987 and 1991, respectively, the M.S. degree in stochastics and operations research from the University of Utrecht, Utrecht, The Netherlands, in 1997, and the Ph.D. degree in statistics from Florida State University, Tallahassee, in 2002.

He is currently an Assistant Professor in the Department of Mathematics and Statistics, Georgia State University, Atlanta. He is the author or coauthor of more than 20 journal papers and conference proceedings. His current research interests include biostatistics, survival analysis, bioinformatics, data mining, machine learning, statistical learning, soft computing, and granular computing.

Dr. Zhao has served as a Program Committee Member for a number of international conferences and is a Reviewer for many international journals and conferences. He was the recipient of the Young Investigators Grant from the National Security Agency.