World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:7, No:7, 2013

Gene Selection Guided by Feature Interdependence

Hung-Ming Lai, Andreas Albrecht, and Kathleen Steinhöfel

***Abstract*—**Cancers could normally be marked by a number of differentially expressed genes which show enormous potential as biomarkers for a certain disease. Recent years, cancer classification based on the investigation of gene expression profiles derived by high-throughput microarrays has widely been used. The selection of discriminative genes is, therefore, an essential preprocess step in carcinogenesis studies. In this paper, we have proposed a novel gene selector using information-theoretic measures for biological discovery. This multivariate filter is a four-stage framework through the analyses of feature relevance, feature interdependence, feature redundancy-dependence and subset rankings, and having been examined on the colon cancer data set. Our experimental result show that the proposed method outperformed other information theorem based filters in all aspect of classification errors and classification performance.

***Keywords*—**Colon cancer, feature interdependence, feature subset selection, gene selection, microarray data analysis.

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

I. Introduction

HIGH-THROUGHPUT screening technologies like microarrays have been widely adopted in recent transcriptome analysis to investigate complete gene expression profiles of cells of interest in response to physiological and genetic changes in many different organisms. Owing to the interrogation of tens of thousands of oligonucleotide probes in parallel, the analysis of the high-throughput data has shown enormous potential for the discovery of biological markers in carcinogenesis studies and in the diagnoses of different types of diseases [1]. By selecting genes with the power of discrimination between cells of normal and tumor, or various tumorigenesis stages, the genome-wide expression-based tumor classification and biomarker discovery derived from these -omics data can be performed. The identification of subsets of these discriminative genes most dedicating to the good predictive classification of cancers is so called gene signatures, and are subject to change [2]. Typically, a small number of signature genes out of the abundance of thousand mRNAs in a tissue sample are very much in favor to establish a final subset, and this is due to significant savings in time and expenses.

The domain of differentially expressed gene selection in bioinformatics is, in fact, an analogy to feature selection that is devised for the need of dimensionality reduction, commonly termed in the context of data mining and machine learning [3]. Feature subset selection techniques aim at reducing the variable dimension of input instances without the change of their initial representation, and searching for the minimal feature subset that maximizes the classification performance or the predictive power. In terms of knowledge discovery, this is actually based on the principle of parsimony [4] that leads to a fact that a model having variables as small as possible to sufficiently fit with the data is preferred – it is exactly what gene signature identification ideally requires. However, the clinical cancer studies using high-throughput biochips merely includes tens to a hundred of samples in their experimental designs and each sample frequently has a large number of questioned features from thousands to tens of thousands of genes [5]. Since feature subset selection is known to be an NP-complete problem [6], the curse of dimensionality would increase the level of the presence of noises that are unavoidable and doomed from the early stages of sample preparation, extraction and hybridization. Microarray-based significant gene selection, therefore, faces a considerable challenge for the model optimization where the removal of large amounts of irrelevant genes could be accomplished.

There are three categories of feature/gene selection methods over the past few years, depending on their connection with the classification method [3]. They are filter, wrapper and embedded techniques. Filter approaches are independent of the classification algorithm and evaluate features relevant to phenotype classes only by interpreting the intrinsic characteristics of the data within an experiment. Filter can be further divided into univariate and multivariate methods. Conventionally, the former covers parametric statistics like paired/unpaired student t-test & ANOVA and nonparametric statistical tests like Wilcoxon rank sum, and considering each feature independently (i.e. neglect of feature interaction). Feature correlations are taken into account to some extent in the multivariate filters which are sometimes referred to as space search filters [7]. Wrapper approaches assess the classification performance or the prediction error of selected feature subsets using a base classifier and search procedures are carried out until satisfactory accuracy is reached. The wrapper usually has a high risk of overfitting and a need of heavy computational demands; on the other hand, it is capable of coping with feature dependence and of interacting with the classifier during the search of feature subset. There are two subclasses of the wrapper, deterministic and randomized methods on the basis of the nature of search algorithms. Typical examples of the former are sequential forward selection and backward elimination while simulated annealing and genetic algorithms dominate the latter [8], [9]. The embedded approach could be viewed as a variant of the wrapper, and having similar properties. The major difference is that feature subset selection is built into the classifier constructions so the embedded method is much less

H.-M. Lai and K. Steinhöfel are with the Department of Informatics, King's College London, Strand, London WC2R 2LS, United Kingdom (e-mail: hung-ming.lai@kcl.ac.uk, kathleen.steinhofel@kcl.ac.uk).

A. Albrecht is with the School of Science and Technology, Middlesex University, Burroughs, London NW4 4BT, United Kingdom (e-mail: A.Albrecht@mdx.ac.uk).

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

computationally expensive than the wrapper. The most popular embedded feature selection is known as SVM-RFE and its variants [10]-[12], where the feature rankings were obtained from the weight vector of a hyperplane using a linear SVM and a preferred feature subset was constructed by a process of recursive feature elimination.

Recently, various feature selection approaches based on information theory have been proposed to measure the correlation among features and the selected feature subset. Here, we briefly introduce three popular space search filters frequently discussed in the research community and will be compared with our new filtered-based gene selector. They all are information theory oriented multivariate methods and seek for the optimal feature subset by taking account of both feature-phenotype correlation and feature-feature dependency. Ding and Peng [13] proposed the minimum-Redundancy Maximum-Relevance framework (mRMR) to reduce mutual redundancy within the feature subset that could capture broader characteristics of classes. Mutual information was exploited in mRMR framework and experimental results showed that their criterion using a combination of relevance and redundancy could lead features to the least redundancy. Fleuret [14] proposed the CMIM feature selection filter and used conditional mutual information as its evaluation criterion to select features maximizing their mutual information with the class given the already selected feature set. Although CMIM claimed that the selected features are not only informative but also weakly pairwise dependency, more potential features might not be captured by the most informative selected features since their procedure would select features having more information about the target class repeatedly. The third feature selector is called FCBF designed by Yu and Liu [15] to efficiently remove large amounts of irrelevant and redundant features. Unlike mRMR and CMIM, FCBF did not introduce an evaluation criterion but incorporated symmetrical uncertainty into an approximate Markov blanket for the feature removal. Thus, FCBF normally chooses fewer features than the other two and it certainly tends to discard some less relevant but important features that might interest the domain expert. In brief, these information theoretic based feature selectors provide a low-dimensional approximation with the consideration of feature relevance and feature redundancy. However, feature interdependence has been neglected in favor of recent feature selection methods. This attribute may sometimes lead to an important feature that are strongly discriminating when accompanied by other features even though the feature is individually weakly relevant – this is especially meaningful to post-genomic gene selection.

In this article, we proposed a multivariate gene selection method guided by interdependent feature pairs that could seem strongly relevant to cancer cell lines or phenotypes. This new gene filter is named iRDA and it provides a four-step framework based on information theory and k-mean clustering. While the calculation of symmetrical uncertainty measures the feature relevance, the devices of k-mean clustering and joint symmetrical uncertainty quantify the strong relevance of feature pairs leading to potentially significant interdependence among features within a selected feature subset. The proposed filter also contains redundant and dependent feature analysis using conditional mutual information and an approach of low-dimensional approximation. Finally, iRDA is able to generate multiple gene signatures and three-way mutual information is applied to ranking these gene subsets for the high level analysis of clinical cancer-related findings. It was observed that our gene selector could find a small number of genes with multiple gene signatures and have remarkable classification accuracy, performance and biological findings on a microarray based gene expression profiling data set.

II. Preliminaries

A. Minimum Feature Selection for Gene Expression Data

A gene expression dataset D from high-throughput technologies can normally be depicted as the following representation. Let $D = \{X \in \mathbb{R}^m, C \in \mathbb{R}\} = \{(x_i, c_i)\}_{i=1}^n$ denote a training set of n labeled samples, where x_i is a sample vector in the feature (gene) space $\mathbb{R}^m$, i.e.$x = \{x_{i1}, \ldots, x_{im}\}$, and C is a class vector to label each extracted sample in an experiment. In a word, the input data of finding minimal significant genes for clinically cancer classification typically consists of a sample matrix and a label vector, illustrating in Figs. 1 (a) and (b). To address the issue of distinctive gene identification at the expression level, we can refer to it as feature subset selection problem. Given $F = \{f_i\}_{i=1}^m$, feature subset selection aims to choose the minimal feature subset $G \subset F$ that maximizes the classification accuracy;namely being an output of gene selection from the dataset D,G has the most discriminative power with respect to the class variable as Fig. 1 (c).

(a) (b)

	Gene 1	Gene 2	Gene 3	·····	Gene m-1	Gene m	Class
Sample 1	2036.28	2253.36	2490.87	·····	1015.91	1459.10	1
Sample 2	1618.65	1066.84	1006.21	·····	853.443	734.529	-1
Sample 3	1597.28	1144.69	1139.63	·····	792.214	1133.24	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sample n-1	4143.38	7256.52	7761.75	·····	2472.23	4906.45	-1
Sample n	2863.18	3036.59	3695.51	·····	1533.80	3030.32	1

(c) {Gene 2, Gene 4, Gene 18, Gene 33, Gene 60}

Fig. 1 Cancer classification using high-throughput screening technologies (a) An example of gene expression profiling data. The experimental dataset contains n samples and each sample has m interrogated genes (m>>n) (b) The extracted samples of an experiment are labeled according to their phenotypes or different types of cell lines. Both a gene expression matrix and a class vector form the inputs of gene selection for cancer classification (c) A subset G of significant genes is obtained as an output for a certain cancer classification, which is so called a gene signature

B. Information Theory

Being an intuitive quantity, entropy is the basic concept of information theory to measure the uncertainty of a random variable and is not calculated according to the real value but at the probability distribution of the variable instead [16]. Given a

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

nominal random variable X, *entropy* is defined based on the representation of Shannon entropy as below

$$\mathrm{H(X)} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \quad (1)$$

where *x* denote possible nominal values of the random variable described in its alphabet $\mathcal{X}$ and $p(x)$ is the marginal probability distribution of X. Without loss of generality, the alphabet of a random variable won't be referred to any longer in this article. The benefit of entropy-based measures is that no prior assumption to be satisfied such as whether data is normally distributed. The information amounts of other events can also be considered by applying the notion of probability theory to information-theoretic entropy. The *conditional entropy* of X given Y is represented as

$$\mathrm{H(X|Y)} = -\sum p(y)\sum p(x|y)\log p(x|y) \quad (2)$$

where $p(x|y)$ is the conditional probability of X when the observed values of Y are given. One could conceive that this measure quantifies the amount of remaining uncertainty of X after the result of Y has been learnt. Similarly, the *joint entropy* of two random variables X and Y is denoted as

$$\mathrm{H(X,Y)} = -\sum\sum p(x,y)\log p(x,y) \quad (3)$$

where $p(x,y)$ is the joint probability distribution of the two variables. This is useful to quantify the uncertainties when two variables should be considered together. Having these entropies defined, we could introduce an essential measure to represent the amount of information shared by two random variables X and Y, *mutual information*, defined by

$$\mathrm{MI(X,Y)} = \mathrm{H(X)} - \mathrm{H(X|Y)} = \sum\sum p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (4)$$

One can regard mutual information as the information quantity that one variable supplies about the other so that this measure is symmetric. Additionally, the value of mutual information is zero, implying that the two variables are statistically independent. Like entropy, the mutual information can be conditioned between X and Y given Z, *conditional mutual information*, and its definition is as follows

$$\mathrm{CMI(X,Y|Z)} = \mathrm{H(X|Z)} - \mathrm{H(X|Y,Z)} = \sum p(z)\sum\sum p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}. \quad (5)$$

The quantity represents the information shared between X and Y after Z is known. Besides the above measures, there is another important quantity in information theory, called *symmetrical uncertainty* that could be seen as a sort of normalized mutual information, defined as

$$\mathrm{SU(X,Y)} = 2\left[\frac{\mathrm{H(X)}-\mathrm{H(X|Y)}}{\mathrm{H(X)}+\mathrm{H(Y)}}\right]. \quad (6)$$

It should be noted that joint symmetrical uncertainty can also be computed if X is a joint random variable in the same way as joint entropy.

C. Feature Relevance

Based on conditional probability distribution, features can be classified as strong relevance, weak relevance and irrelevance proposed and defined by Kohavi and John[17]. Given a full set of features F and a feature f_i, the subset of features $F_i = F\backslash f_i$ denotes a full set of features excluding feature i. The following feature definitions (Def 1-3) have been introduced by Kohavi and John.

Definition 1: Strongly Relevant Feature
A feature f_i is strongly relevant to C iff

$$p(\mathrm{C}|f_i, F_i) \neq p(C|F_i). \quad (7)$$

Definition 2: Weakly Relevant Feature
A feature f_i is weakly relevant to C iff

$$p(\mathrm{C}|f_i, F_i) = p(C|F_i) \text{ and}$$

$$\exists F_i' \subset F_i \text{ such that } p(\mathrm{C}|f_i, F_i') \neq p(C|F_i'). \quad (8)$$

Definition 3: Irrelevant Feature
A feature f_i is irrelevant to C iff

$$\forall F_i' \subseteq F_i,\ p(\mathrm{C}|f_i, F_i') = p(C|F_i'). \quad (9)$$

The above definitions could imply that it is compulsory to always include strongly relevant features and some weakly relevant features that are not redundant to one another in an optimal feature subset, and irrelevant features are certainly excluded. Similar to these considerations over a single feature, one can define the strong relevance of a feature pair i and j using joint random variables $f_i f_j$ (or sometimes denoted as f_{ij}).

Definition 4: Strongly Relevant Feature Pair
A feature pair f_{ij} is strongly relevant to C iff

$$p(\mathrm{C}|f_{ij}, F_{ij}) \neq p(C|F_{ij}) \quad (10)$$

where F_{ij} is a full feature set that has excluded feature i and j, and this means that one can view a feature pair as a united-individual when selecting features. The notion of strongly relevant feature pairs will be fundamental to the proposed filter-based gene selection where feature pairs with strong relevant to the target class would be included as many as possible in the selected procedure.

III. iRDA– A Novel Gene Selection Filter

A new feature selection filter named interdependence-guided gene selection with redundant and dependent analysis (iRDA), shown in Algorithm 1, is proposed for high-throughput screening gene selection. This method is based on information-theoretic measures including symmetrical

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

uncertainty (SU), mutual information (MI) and conditional mutual information (CMI) to find potential gene subsets which has the most power of discrimination for cancer cell classification and biomarker discovery. As mentioned in preliminaries, the optimal gene subset has to contain all features of strong relevance and some ones of weak relevance. A gene can hardly ever function itself, but co-regulation between genes is always the case in a cell line instead. Interdependence between features is, therefore, a matter when it comes to significant gene subsets.

The proposed gene selector is a four-stage filter that contains the analyses of feature relevance, feature interdependence, redundancy-dependence, and subset rankings. In practice, it is quite impossible to make use of Kohavi and John's definitions to find and filter useful features, in particular in large-scale feature selection. In order to examine features for their extent of relevance with the target classes, symmetrical uncertainty of a random variable (feature f_i) with respect to given labels is calculated for each feature, $SU_{i,c}$ (line 2). These correlations are first sorted in descending order and then k-mean clustering is proceeded on the sorted list to disclose the degree of feature relevance. Due to the characteristic of relatively small number of expressed genes in large scale gene expression profiles, the sorted correlations are actually exponentially distributed. Thus, five clusters are chosen to draw up different scales in terms of relevance, representing Very Strong (R1), Strong (R2), Weak (R3), Very Weak (R4) and Irrelevant (R5) feature types (lines 3-5). The drawn scales are a perquisite of performing primary idea of iRDA, interdependent feature analysis.

Although high-order gene interactions are ideally targeted, it would often result in a disaster of a larger number of time-consuming calculations. Only is interdependence between two features considered here, and we utilize symmetrical uncertainty of a joint random variable of two features $f_i f_j$ with respect to target classes to gain their correlations for any two features, $SU_{ij,c}$. The aim in the step of interdependent analysis is to seek for strongly relevant feature pairs whose joint SU values are greater than a thresholdε (line 11). Those features in the R1 cluster colliding with another feature in the clusters of R1, R2, R3 and R4, with a view to tackling high dimensional two-way interdependence, are regarded as candidate pairs (lines 7-16). Through the scheme of relevance partition, the value of ε is also easily to be estimated – the mean of joint SU values of the most strongly relevant feature in the very strong cluster crashing with the other features in the clustersR1-R4, the irrelevant cluster excluded, i.e. $\overline{SU_{1J,c}}$. The successful feature pairs are then added into a subset of G_t guided by a common feature, called a kernel, among feature pairs in order of relevance. It should be noted that the kernel feature is added once in the same subset.

ALGORITHM 1. iRDA Filter

Given: $D = \{X \in \mathbb{R}^m, C \in \mathbb{R}\} = \{(x_i, c_i)\}_{i=1}^n$ and $F = \{f_i\}_{i=1}^m$
Parameters: k, ε
Find: gene signatures *G*

1 **Relevance:**
2 for each f_i in *F*calculate$SU_{i,c}$ and sort them in the descending order
3 Perform k-mean clustering (k=5) on the list of the sorted $SU_{i,c}$, and tag
4 them in sequence, Very Strong(R1), Strong(R2), Weak (R3), Very Weak
5 (R4) and Irrelevance (R5).
6 **Interdependence:**
7 t=1
8 for i=1 to sup(R1)
9 $G_t = \emptyset$
10 for j=i+1 to sup(R4)
11 if $SU_{ij,c}$>ε, where ε is estimated by $\overline{SU_{1J,c}}$
12 add feature pairs $f_i f_j$ into G_t in order of relevance,
13 where f_i is a kernel and added once
14 end
15 t=t+1
16 end
17 **Redundancy and Dependence:**
18 Let $G_{pre} = \{G_t | G_t \neq \emptyset\}$, $\forall G_t \in G_{pre}$, G_t is kernel-guided
19 for each G_t do
20 for each $f_i \in G_t$ in a less relevant sequence do
21 f_i is removed instantly if $CMI(f_i; C|G_t)$=0
22 $G_t = \emptyset$ if $CMI(kernel; C|G_t)$=0
23 end
24 end
25 for each$f_{kernel,j}$ in G_{pre}, add f_{kernel} into the subset guided by f_j if
26 applicable
27 perform lines 19-24 again
28 Let $G_{post} = \{G_t | G_t \neq \emptyset, \#G_t > 1\}$
29 **Subset Rankings:**
30 Rank each G_t in G_{post} using (11), and top K gene signatures are selected
31 to establish *G*

As soon as feature pairs are gained from the second stage, features can be able to line up in a row after a common feature (kernel) guiding other interdependent features to form a subset, G_t. A set of these feature subsets, G_{pre}, would then be carried on the step of redundant and dependent analysis (line 18). For all possible subsets, conditional mutual information of a feature f_i and labels C given a selected subset $G_t \in G_{pre}$, CMI($f_i; C|G_t$), is exploited to measure if a feature is redundant in its residential subset (lines 19-24). A backward elimination and less relevant feature first policy is the main approach of redundancy approximation during this stage. A feature with the lower value of $SU_{i,c}$, i.e. less relevance, is first examined and if its CMI value is zero, the feature is removed instantly and followed by checking the next less relevant feature; the procedure goes on until all features in the same subset are checked. If a kernel is removed, the subset guided by the kernel will be eliminated; otherwise features remain in a retained

subset will be viewed as dependents of the kernel. Also, for each kernel-connected feature pairs, $f_{kernel,j}$ in G_{pre}, we may add the kernel feature into the subset guided by feature j if applicable. Following this post-process of subset formation, iRDA need to perform the second round of CMI examination and the potential subsets would then be collected (lines 25-28). Briefly speaking, through the mechanism of dependents following after the same interdependence-guided gene, interdependence between two genes could reveal more sophisticated interaction among multiple genes when a subset survives the redundant-dependent stage. A set consisting of these subset survivals, G_{post}, finally go forward to the last step where ranking is applied to each candidate subset (lines 30-31).

Since our method does not target on individual genes but gene subsets instead, subset rankings are the issue and cell classification can be predicted not only by one of top-ranking subsets but also by an ensemble means across multiple subsets. The latter is in particular suitable for large scale gene selection with small sample size. Mutual information of three random variables (kernel, subset and labels) is used to measure the subset ranking and low-dimensional approximation of this joint MI is also devised as the following equation.

$$MI(kernel, G_t; C) = \frac{1}{\#G_t}\sum\sum MI(f_i f_j; C) + MI(f_k; C) + \frac{1}{\#G_t}\sum dep(f_k, f_j) \quad (11)$$

where $dep(f_k, f_j) = [CMI(f_k, f_j|C) - MI(f_k, f_j)]$ /h (f_k, f_j), f_k: kernel feature, $h(f_k)$: entropy, h(f_k, f_j): joint entropy, $\#G_t$: cardinality of G_t.

When subset rankings are performed, domain users can select top K subsets as candidates of biological interests. To illustrate how biological discovery might be found using multiple gene signatures, we exploit first three ranks of subsets and then classification could be predicted by an ensemble means across the three rankings (say S_1, S_2 and S_3). In general, majority voting is our approach. We would accept minority of classification if likelihood of minority was greater than the maximum of likelihood of majority; otherwise the consequence of majority voting would be accepted.

IV. Experimental Results

To demonstrate the effectiveness of the novel gene selection filter, we selected a publicly available microarray-based colon cancer data set. This data set was introduced by Alon et al. [18] and frequently used to validate the performance of cancer classification and gene selection in the research community. The colon cancer data set consists of 62 samples from the patients of colorectal cancer (CRC), where 22 normal labels are extracted from healthy tissues and 40 abnormal biopsies are extracted from colon tumors. Out of more than 6,500 genes in the original design of experiment, 2,000 genes were selected by the authors for clustering functional genes and classifying tissues. In this article, the experiment of colon cancer gene selection was performed in the environment of Matlab 7.14 with the third party tool, MIToolbox [19], upon the computer of Intel Core i5 with 2.50 GHz processor and 8GB RAM. For the better calculation of information theoretic measures, features were discretized to three bins as suggested by Ding and Peng [13] and each bin was then designated by a discrete value such as 1, 3 and 5.

TABLE I
Feature Selectors on Crc data Set

Filter	Error	AUC	Genes	Authors
mRMR	13/10	0.8256/0.8199	11/2	Ding andPeng[13]
CMIM	12/10	0.8398/0.8432	11/4	Fleuret[14]
FCBF	9	0.8778	9	Yu and Liu[15]
iRDA	4	0.9176	11	

LOOCV gene selection was performed on 3NN classifier

Working on the CRC data set, we evaluated our method, iRDA, with three model-free feature filters of mRMR, CMIM and FCBF to know the characteristics of the proposed gene selector in terms of classification error and performance. Because of the curse of dimensionality, the conventional training-test data partition given a ratio (say 60-40%) is not very appropriate for the assessment of gene selection approaches in the domain of high-throughput gene expression data. In the study, the procedure of leave-one-out cross-validation (LOOCV) was used to assess the performance of selectors and the classification error estimation during a selection procedure. Additionally, a reference classifier is also needed to induct the selected feature selectors into a learning process. This is due to their independence of learning methods. We utilized the k-nearest-neighbor (k-NN) classifier (k=3) to establish classification models after gene selectors had been performed. The experimental result of the four filters using the colon cancer data set is shown in Table I. In the clinical cancer study, not only false-positive (FP) but also false-negative (FN) errors are vital to be estimated since they have different operating characteristics for cancer classification. Here, we simply calculated FP and FN as classification errors.

To understand the classification performance of different feature selection methods, we employed the area under a Receiver Operating Characteristic (ROC) curve, abbreviated as AUC. AUC is able to summarize the ROC curve which is a plot of the sensitivity against 1 –specificity to visualize the power of a classifier. Since iRDA is a multiple gene signatures oriented selector, the proposed method in general identified two more genes than FCBF where nine genes or so are maximally filtered. Apart from the best results of the four filters, their performances and classification errors at the level of the same/similar number of signature genes were reported. Table I shows that iRDA had four misclassifications and more than 90% classification performance at the AUC level while the other three had about ten classification errors and their prediction performances were from 82% to 88%, much worse than that of iRDA. The minimal signature genes of mRMR and CMIM were quite few (2 and 4 respectively), but their predictive power would be viewed as a moderate degree of success (82% and 84% respectively) even if they increased the number of selected genes up to fifty. Although the significant genes found by iRDA were slightly more than those filtered by FCBF, it seems still reasonable to have the scale of the selection. In sum, one can easily find that iRDA is by far the

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

most promising gene selector among them on the study of CRC data set.

TABLE II
SUMMARY OF GENE SIGNATURES

Set	Gene Name
1	HSPD1, SPARC, S100P, ZNF117
2	MYL2, EIF2S2, RPL24, A2M
3	SPARC, HSPD1, CDH3, HIVEP2, PSMC2

By applying the iRDA gene selector into the whole sample sets of previous CRC data, we can obtain multiple colon cancer associated gene signatures and the first three subsets ranked by iRDA are summarized in Table II. These could be candidates of clinical interest to the domain expert. To illustrate how well the novel gene selector operates in practice from biological point of view, one of gene signatures identified by iRDA was studied through literatures and KEGG signaling pathway. The signature contains four genes of MYL2, EIF2S2, RPL24 and A2M. From the queries of KEGG, We found that MYL2 and A2M belong to the pathway of immune system that is a subclass of organismal systems and both EIF2S2 and RPL24 participate in the translation process of genetic information processing. We noticed that A2M was filtered only by iRDA so it might be interesting to know if this gene is biologically significant to the CRC development. Ghilardi et al. reported that MMP-7-181 G/G genotype is involved in colon cancer and tumor progression in Italians [20] while Yang et al. claimed that MMP-7 also plays an important role in tumor development and progression process in CRC patients in China [21]. And a recent report showed that the main inhibitor of MMPs in tissue fluids is α2-macroglobulin (A2M) [22]. On the other hand, we also noticed that one of significant genes, named HSPD1, was filtered by all of the four selectors and most recent research summarized that HSP60 (i.e. HSPD1) could be a potential biomarker in CRC [23]. Thus, it is believed that the proposed method could have a high capacity of finding the majority of significant genes that most filters would also identify and of finding other important genes involved in gene regulation that most selectors could ignore.

V. CONCLUSIONS AND FUTURE RESEARCH

The proposed filter-based gene selector has successfully been tested on one of typical clinical cancer classification experiments, colon cancer data set. By using this data set, we have compared iRDA with three filters (mRMR, CMIM and FCBF) that are widely used and also based on information theorem. Our experimental results show that iRDA outperformed them in all aspects of classification errors and classification performance and could have great potential for biomarker discovery. Large dimensional variables with small sample size would cause that the optimal gene subset could not be a unique. iRDA provides multiple feature subsets with different rankings could be more suitable for large-scale gene expression profiling analysis. This is very different from conventional filter methods where various K genes are selected and form a subset for high level gene analysis. In addition to having multiple gene signatures, the new filter method is able to find an important feature that is individually weakly relevant but has strong interdependence between features. This type of genes accompanied by other significant genes would more contribute to the phenotype than they appear solely at the expression level. Unfortunately, most recent filter-based feature selectors could not search for these features that may attract the interest of the domain user. Although an early microarray cancer data set was performed for the effectiveness of the proposed gene selector, we could still reveal several significantly biological findings on the colorectal cancer research or discover potential biomarkers using the iRDA filter with recent literature studies and signal pathway databases.

There are a couple of issues that will be addressed for our future research. Experimental mRNA expression data is cursed by its nature of small sample size with large dimensional features. The unavoidable problem always leads to unstable gene signatures which has recently attracted many researchers' attention. Being a developer of gene selectors, we will have to deal with this matter to see how robust the proposed method could be and whether any skills could be well incorporated into it if the developed selector is not stable enough. Although the effectiveness of iRDA has been successfully validated on the colon cancer data set, most recent gene expression profiling data for other tumor-associated classification are required. Most cancer classification data sets, like Alon's CRC data set, were introduced at an early age of microarray where better preprocessing techniques and experimental designs were still limited, and the two events also have a high impact on the removal of experimental noises. Therefore, it is worth connecting iRDA with up-to-date cancer-related classification data generated from high-throughput technologies to further understand the characteristics of the proposed gene selector.

REFERENCES

[1] J. R. Nevins, and A. Potti, "Mining gene expression profiles: expression signatures as cancer phenotypes," *Nature Reviews Genetics*, vol. 8, no. 8, pp. 601-609, Aug, 2007.
[2] S. Y. Kim, "Effects of sample size on robustness and prediction accuracy of a prognostic gene signature," *BMC Bioinformatics*, vol. 10, pp. 147, May, 2009.
[3] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, Oct, 2007.
[4] D. A. Bell, and H. Wang, "A formalism for relevance and its application in feature subset selection," *Machine Learning*, vol. 41, no. 2, pp. 175-195, Nov, 2004.
[5] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings National Academy Sciences*, vol. 103, no. 15, pp. 5923-5928, Apr, 2006.
[6] S. Davies, and S. Russell, "NP-completeness of searches for smallest possible feature sets," *AAAI Symposium on Intelligent Relevance*, pp. 37-39, 1994.
[7] C. Lazar, J. Taminau, S. Meganck*et al*., "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106-1119, Jul-Aug, 2012.
[8] A. Albrechta, S. A. Vinterbob, and L. Ohno-Machado, "An Epicurean learning approach to gene-expression data classification," *Artificial Intelligence in Medicine*, vol. 28, no. 1, pp. 75-87, May, 2003.

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf

[9] I. A. Gheyas, and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, no. 1, pp. 5-13, Jan, 2010.
[10] I. Guyon, J. Weston, S. Barnhill *et al*., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389-422, 2002.
[11] X. Zhou, and D. P. Tuck, "MSVM-RFE:extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106-1114, May, 2007.
[12] P. A. Mundra, and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans Nanobioscience*, vol. 9, no. 1, pp. 31-37, Mar, 2010.
[13] C. Ding, and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185-205, Apr, 2005.
[14] F. Fleuret, "Fast binary feature selection with conditional mutualinformation," *Journal of Machine Learning Research*, vol. 5, pp. 1531-1555, Nov, 2004.
[15] L. Yu, and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205-1224, Oct, 2004.
[16] T. M. Cover, and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Hoboken, NJ: John Wiley & Sons, ch. 2, pp. 13-55, 2006.
[17] R. Kohavi, and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, Dec, 1997.
[18] U. Alon, N. Barkai, D. A. Notterman*et al*., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings National Academy Sciences*, vol. 96, no. 12, pp. 6745-6750, Jun, 1999.
[19] G. Brown, A. Pocock, M.-J. Zhao *et al*., "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research*, vol. 13, pp. 27-66, Jan, 2012.
[20] G. Ghilardi, M. L. Biondi, M. Erario*et al*., "Colorectal carcinoma susceptibility and metastases are associated with matrix metalloproteinase-7 promoter polymorphisms.," *Clinical Chemistry*, vol. 49, no. 11, pp. 1940-1942, Nov, 2003.
[21] B. Yang, K. Su, J. Gao*et al*., "Expression and prognostic value of matrix metalloproteinase-7 in colorectal cancer," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 3, pp. 1049-1052, 2012.
[22] M. Egeblad, and Z. Werb, "New functions for the matrix metalloproteinases in cancer progression," *Nature Reviews Cancer*, vol. 2, no. 3, pp. 161-174, Mar, 2002.
[23] Y. Ma, P. Zhang, F. Wang *et al*., "Searching for consistently reported up- and down-regulated biomarkers in colorectal cancer: a systematic review of proteomic studies," *Molecular Biology Reports*, vol. 39, no. 8, pp. 8483-8490, Aug, 2012.

Open Science Index, Biotechnology and Bioengineering Vol:7, No:7, 2013 publications.waset.org/16510.pdf