6-2004

Algorithmic Fusion of Gene Expression Profiling for Diffuse Large B-Cell Lymphoma Outcome Prediction

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Hongmei Cui

Kahai Cao
University of Nebraska at Omaha

Wing C. Chan

Algorithmic Fusion of Gene Expression Profiling for Diffuse Large B-Cell Lymphoma Outcome Prediction

Qiuming Zhu, *Senior Member, IEEE*, Hongmei Cui, Kajia Cao, and Wing C. Chan

***Abstract*—Many different methods and techniques have been investigated for the processing and analysis of microarray gene expression profiling datasets. It is noted that the accuracy and reliability of the results are often dependent on the measurement approaches applied, and no single measurement so far is guaranteed to generate a satisfactory result. In this paper, an algorithmic fusion approach is presented for extracting genes that are predictive to clinical outcomes (survival–fatal) of diffuse large B-cell lymphoma on a set of microarray data for gene expression profiling. The approach integrates a set of measurements from different aspects in terms of the discrepancy indications and merit expectations of the gene expression patterns with respect to the clinical outcomes. A combination of statistical and nonstatistical criteria, continuous and discrete parameterizations, as well as model-based and modeless evaluations is applied in the approach. By integrating these measurements, a set of genes that are indicative to the clinical outcomes are better captured from the gene expression profiling dataset.**

***Index Terms*—Algorithmic fusion, cross-projection (CP), discrete partition (DP), Fisher's discrimination, gene expression profiling, outcome prediction.**

I. Introduction

DNA MICROARRAY technology provides biologists with the ability to measure the expression levels of thousands of genes simultaneously in a single experiment. It is essential to have some effective means to extract the biologically significant information from the large amount of microarray datasets. Identifying genes that are predictive to certain clinical outcomes is one of the important goals in the analysis of gene expression profiling data from tumor specimens [1].

Two gene expression profiling studies of diffuse large B-cell lymphoma (DLBCL) had been undertaken by researchers to identify genes predictive to clinical outcomes [16], [17]. Rosenwald *et al.* identified the functional (gene expression signature) groups of 17 genes that are predictive to survival from a dataset of over 7394 genes on 154 clinical cases using statistics-based evaluations. They found that the groups of genes are closely associated with the genes that divide the tumor into distinct biologic subtypes [16]. Shipp *et al.* applied a supervised learning method on an expression profiling dataset of 7139 genes on 58 tumor specimens, and identified 13 genes that are highly predictive to the outcomes [17]. The research presented in this paper uses Shipp's dataset (www.genome.wi.mit.edu/MPR/lymphoma) and complements Shipp's outcome predictors with our results based on a multimeasurements algorithmic fusion approach.

There are a number of methods discussed in the literature for identifying genes that are indicators of certain diseases or health disorders. General approaches include 1) parametric methods, such as the principal component analysis [13], independent component analysis, and separation correlation metric (SCM), alternatively known as the Fisher's discrimination criterion (FDC) [7]; and 2) nonparametric methods, such as the projection pursuit regression (PPR) [9], support vector machines [4], neural networks, threshold number of misclassification (TNoM) [3], and expectation maximization [6]. Most methods resulted in certain scoring of individual genes for the objective-relevance detection.

The parametric methods use a set of statistical metrics derived from the gene expression profiling under the assumption of certain probability distribution models. Statistical methods are usually reliable and accurate in large data set analysis, especially for the methods that apply the class of robust statistics [10]. However, the incongruence between the relatively small number of data samples collected in current practice and the high dimensions of the genes profiles often makes the statistical model hard to be justified, one version of the so-called dimension curse in scientific computation. Moreover, the statistical measurements are easily biased or distorted by the uncertainty and inaccuracy of the sample labels, the exact category of the specimen, and irregularity of the sample distributions.

The nonparametric methods do not rely on the assumption of statistical models and parameters. Rather, they work directly toward the objectives, such as discrimination or prediction, by applying certain nonstatistical metrics on the data samples. These methods are advantageous at constraining and attenuating the effects of smaller sample numbers. However, the diversity of measurement metrics and the uncertainty, including the imprecision and incompleteness of the data set, of the individual data samples often make it hard to obtain consistent results in different experimentations. That is, the results are easily affected by the experimental situation and environmental conditions. The inherent variability of the cases from the clinical/pathologic diagnoses also affects the feature identification. For example, the survival data in the DLBCL cases are not only determined by the genomic alterations in the tumor, or the treatment applied, but also by the diverse types of individual circumstances such as the patient's age, general health conditions, and specific drug metabolism. This type of inherent variability affects both parametric and nonparametric approaches.

Q. Zhu, H. Cui, and K. Cao are with the Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182 USA (e-mail: zhuq@unomaha.edu).

W. C. Chan is with the Department of Pathology and Microbiology, Medical Center, University of Nebraska, Omaha, NE 68182 USA.

Digital Object Identifier 10.1109/TITB.2004.828894

It is known that measurement of some attributes of a set of objects is the process of assigning numbers or other symbols to the objects in such a way that the relationships of the objects being measured [2], [18]. A particular way of assigning numbers or symbols according to the attributes of the objects is called a scale of measurement [5]. These kinds of problems are the subject of study in measurement theory. Measurement theory shows that strong assumptions are required for certain statistics to provide meaningful information about reality of the object in concern. Often the only thing for sure is that the measurement is a monotone increasing function of valuation. In many cases, people would like to choose an analysis that yields invariant results. The study of such invariants is a major concern of measurement theory. However, measurement theory does not provide a complete solution to such problem. In particular, measurement theory does not take random measurement error into account, and if such errors are an important aspect of the measurement process, then additional analysis methods are called for.

It is intuitive to imagine that a better result could be obtained in general by combining a number of the measurements together [19]. However, the quality of the result relies on how the different measurements are integrated and what is the set of the measurements that form a good combination of measurement for a given set of data and measurement objectives. While a selected set of measurements might be monotonic increasing, some of them are in continuous values and some are in discrete scales. Moreover, some of the measurements are in statistical nature and some are nonparametric, and some of them are model-based and some are modeless as well. These factors pose various kinds of difficulties for the measurement combinations.

In this paper, we present an algorithmic fusion method that is based on an integration of a set of parametric and nonparametric measurements in different scoring/ranking metrics. The method extracts a group of genes that are predictive to the DLBCL clinical outcome by integrating the measurements in a crossover ranking-and-selection process. In the following, the major analytic methods used in the algorithmic fusion approach will be presented in Section II. Section III describes the algorithm for fusing the multiple measurements to generate a unified set of predictive genes. The results obtained from applying the approach to the sample DLBCL data set for outcome prediction are presented in Section IV. Section V presents concluding remarks.

II. Multiple Measurements

The basic idea of measurement theory is that a quantitative scale is a mapping between some objects and a given set of associated numerical values. This mapping, however, is not arbitrary but is supposed to meet some requirements. It is important to be mindful of certain measurement procedures and to recognize fully their strengths and shortcomings [5]. In gene expression profiling, different algorithms and evaluation techniques often result in a different set of candidate genes that are extracted for outcome predictions. It is, therefore, desirable to have a set of criteria established in the analysis so as to control the gene selecting process, and to properly assess the merit of the resulting gene sets from the measurements. We adopted the "distinctiveness" and "representativeness" as our major criteria in this research. That is, the technique discussed in this paper is aimed at extracting the gene set that are as much "distinctive" as possible with respect to different outcome classes so that the distributions of the gene values are well separated, that is, the gene expression levels are numerically distinguishable. These kinds of sets of genes are often called predictors. In the following, we discuss some typical measurement methods that address their special characteristics for revealing the gene expression discrepancies with respect to different clinical outcomes in terms of the above criteria.

A. SCM/FDC as a Class Separation Metric

The SCM, also known as the FDC [7], is a well-known and popular parametric method for identifying data attributes and their projections that are most likely separable among different classes. It has been used in several applications for extracting genes that are differentially expressed diseases or with different clinical outcomes [4]. However, it is also known that the FDC is not an absolute criterion for yielding accurate classifications [20]. The method should be combined with other correlation analyses in practical applications to diminish some parametric side effects.

Let ω_1 and ω_2 be the labels of two different sample classes, for example, the surviving versus fatal cases of the DLBCL. The one-dimensional (1-D) SCM/FDC method aims at maximizing the criterion

$$J(\underline{W}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{\underline{W}^T S_B \underline{W}}{\underline{W}^T S_W \underline{W}}$$

where $\mu_i.$, $i = 1, 2$, is the mean value of the 1-D projection of the data samples of classes ω_i in a direction expressed by a vector $\underline{W}$, respectively. That is, $\mu_i = (1/n_i) \sum_{\underline{X} \varepsilon \omega_i} \underline{W}^T \underline{X}$, where $\underline{X} = [x_1, x_2, \ldots]$ represents the gene expression vector which consists of expression values of the individual gene over all clinical samples. The n_i is the number of data samples in class ω_i. The $\sigma_i^2 = (1/n_i) \sum_{\underline{X} \in \omega_i} (\underline{W}^T \underline{X} - \mu_i)^2$, $i = 1, 2$, is the variance for the projected samples of class ω_i in direction $\underline{W}$, respectively. $\underline{W} = [w_1, w_2, \ldots, w_n]$ is a vector in the $\underline{X}$ space that serves as a transformation operator on which the samples X are projected to a 1-D space $y = \underline{W}^T \underline{X}$. $S_B = (\underline{m}_1 - \underline{m}_2)(\underline{m}_1 - \underline{m}_2)^T$ is called the between class scatter matrix, where $m_i.$, $i = 1, 2$, is the mean vector of the data samples of classes ω_I in the original vector space of $\underline{X}$. $S_W = S_1 + S_2$ is called the within-class scatter matrix, where S_i, $i = 1, 2$, is the covariance matrix of the data samples of classes ω_i in the space of $\underline{X}$.

When limiting the projection vector $\underline{W}$ to the form of $[1\ 0 \ldots]$, $[0\ 1 \ldots]$, $\ldots$ that is, the axes of the Euclidean coordinates, the 1-D SCM/FDC represents a measurement of individual gene according to its mean and variance parameters with respect to the original class designations. Let FDC_k denote such a measurement on gene g_k, i.e, on vector $x_k = [x_{k1}, x_{k2}, \ldots]$, the SCM/FDC criterion (denoted as FDC simply) can be expressed as

$$\mathrm{FDC}_k = \frac{(\mu_{k1} - \mu_{k2})^2}{\frac{\left(\sum_{j=1}^{n_1} \left| x_{kj}^{(1)} - \mu_{k1} \right|\right)}{n_1} + \frac{\left(\sum_{j=1}^{n_2} \left| x_{kj}^{(2)} - \mu_{k2} \right|\right)}{n_2}}$$

where the $x_{kj}^{(1)}$ and $x_{kj}^{(2)}$ are the kth gene expression values of the jth sample corresponding to classes ω_1 and ω_2, respectively.

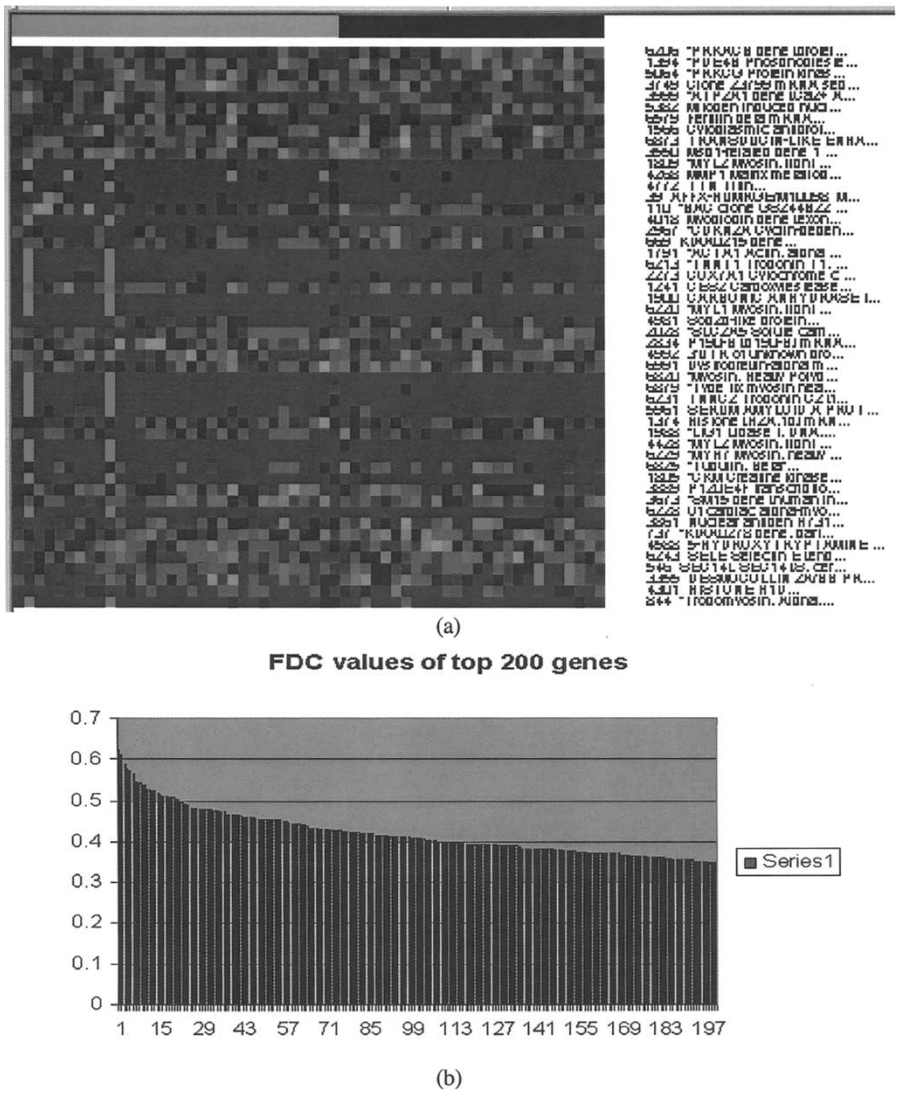

Fig. 1. FDC measurement of the DLBCL dataset. (a) Genes of top 50 FDC ranks. (b) FDC values of top 200 genes.

The FDC values provide a means for ranking the separability of the genes with respect to the tumor outcome classes. Our experimental result of the FDC measurements on the DLBCL dataset is illustrated in Fig. 1, where Fig. 1(a) presents the genes ranked at the top 50 in the FDC measurement and Fig. 1(b) shows the FDC values of the top 200 genes. The top bar with color green and black in Fig. 1(a) indicates the sample cases in survival and fatal outcomes, respectively. From Fig. 1, we can see that the FDC measurement does not provide an overall good indication of the genes that are related to the survival–fatal outcomes. A number of genes even ranked at the top ten of relatively high FDC values do not show a clear pattern of separation for the samples in two different outcomes. That is, not every gene with the high FDC value is distinctive toward the separation of two distinct outcome classes.

B. Cross-Projection (CP) Index as a Maximum Likelihood Measurement

The generic term "projection pursuit" refers to two statistical procedures: exploratory projection pursuit and PPR. Friedman and Tukey coined the term projection pursuit for a technique of finding interesting low-dimensional projections of high-dimensional data set [9]. Since then, projection pursuit has been a general method for exploratory data analysis. Fridman also characterized a given projection by a numerical index that can then be used as a basis of a heuristic search to locate the "interesting" projections [8]. A projection pursuit index concerns with associating a functional value to each and every low-dimensional projection that reveal the most details about the structure of the data set [11]. Once an interesting set of projections has been found, the existing structures (e.g., a set of genes, data clusters, etc.) of different patterns can be extracted and analyzed separately by the indexes [15]. Usually this index value is derived from an evaluation function associated with the projection. The function itself is called the projection indexing and is usually differentiable to facilitate efficient optimization [12]. A search is usually conducted for revealing correlative projections by maximizing the index over the projection space, that is all possible projections of the function.

To identify genes that are predictive of surviving and fatal DLBCL tumor cases in terms of maximizing the likelihood of

classification, we applied the projection pursuit method to the gene expression profiles. Our method projects the gene expression values of individual case sample to the corresponding likelihood function with respect to the distribution of the whole class. We call it a CP. The quantitative measurements of the projections over all cases are accumulated to form the index. As the projections take place between different pairs of gene patterns or class spaces, a number of genes reflecting the characteristics of the gene expressions among different classes are extracted according to the scoring of the projection indexing values.

A CP process and the computation of the cross-projection index (CPI) take the following steps in general.

1) Considering two data sets D_1 and D_2, where

$$D_1 = \{x_1^{(1)}, x_2^{(1)}, \ldots, x_{n1}^{(1)}\}, \text{ and } D_2 = \{x_1^{(2)}, x_2^{(2)}, \ldots, x_{n2}^{(2)}\}.$$

Data points $x_k^{(1)} \in D_1$, $k = 1, 2, \ldots, n_1$ are case samples from a class ω_1.
Data points $x_k^{(2)} \in D_2$, $k = 1, 2, \ldots, n_2$ are case samples from a class ω_2.
Each point $x_k^{(c)}$, $c \in \{1, 2\}$ is a vector $x_k^{(c)} = [x_{k1}^{(c)}, x_{k2}^{(c)}, \ldots, x_{km}^{(c)}]$, where m is the number of data attributes (genes in our experiments) in each case sample.

2) Assuming that the elements of the data point are in Gaussian distributions

$$\begin{aligned} p^{(1)}(x) &= p(x|x \in \omega_1) = p\left(x_{kj}^{(1)}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2} \\ p^{(2)}(x) &= p(x|x \in \omega_2) = p\left(x_{kj}^{(2)}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2} \end{aligned}$$

where $\mu_j^{(c)} = (1/n_c)\sum_{k=1}^{n_c} x_{kj}^{(c)}$, and $\sigma_j^{(c)} = (1/n_c)\sum_{k=1}^{n_c}(x_{kj}^{(c)} - \mu_j^{(c)})^2$, $c \in \{1, 2\}$; and n_c is the number of data points in class ω_c.

3) A CP of $x_{kj}^{(1)}$, the jth element of data point x_k of class ω_1, to the distribution of class ω_2 is defined as

$$p^{(2)}\left(x_{kj}^{(1)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2}.$$

Similarly, we have the CP of $x_{kj}^{(2)}$, the jth element of data point x_k of class ω_2, to the distribution of class ω_1

$$p^{(1)}\left(x_{kj}^{(2)}\right) = \frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} e^{-\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2}.$$

4) Taking a logarithm transformation of the above expressions, we have

$$\begin{aligned} P^{11} &= \log\left(p^{(1)}\left(x_{kj}^{(1)}\right)\right) \\ &= -\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} \\ P^{12} &= \log\left(p^{(1)}\left(x_{kj}^{(2)}\right)\right) \\ &= -\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(1)}}{\sigma_j^{(1)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}} \\ P^{21} &= \log\left(p^{(2)}\left(x_{kj}^{(1)}\right)\right) \\ &= -\frac{1}{2}\left(\frac{x_{kj}^{(1)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(2)}} \\ P^{22} &= \log\left(p^{(2)}\left(x_{kj}^{(2)}\right)\right) \\ &= -\frac{1}{2}\left(\frac{x_{kj}^{(2)} - \mu_j^{(2)}}{\sigma_j^{(2)}}\right)^2 + \ln\frac{1}{\sqrt{2\pi}\sigma_j^{(2)}}. \end{aligned}$$

These measurements are the likelihood of the sample vectors with respect to different outcome classes under inspection.

5) The CPI is computed as follows:
Let
$Q_j^{(1)}$ be the number of data points $x_k^{(1)}$ on gene j such that $[P^{11} > P^{21}]$, and
$Q_j^{(2)}$ be the number of data points $x_k^{(2)}$ on gene j such that $[P^{22} > P^{12}]$.
A CPI of gene g_j is defined as

$$\text{CPI}(j) = Q_j^{(1)} + Q_j^{(2)}.$$

A graphical display of the expression profiles for 50 genes with high CPI values in the DLBCL dataset is shown in Fig. 2(a). The CPI values of the top 200 genes are shown in Fig. 2(b). Note that the CPI measurements are discrete. It is a scoring of the maximum likelihood of a data point x_k with respect to a class ω_i. A continuously valued CPI can also be defined such that it takes the count of accumulative differences of the likelihood values on the CPs. It yields the similar results as the CPI defined above.

C. DP (Discrete Partition) Index as a Minimum Misclassification Measurement

While the FDC and CPI are parametric methods for data evaluation, we applied a nonparametric and discrete method for a measurement of the minimum misclassification (error) rate on the data set. The technique is based on the principle of the TNoM [3]. The TNoM approach tries to find a data point V_k on a gene expression profile over all test cases such that the total misclassified samples would be minimized if the data point V_k is taken as a separation threshold to divide the gene expression values into two distinct classes. The minimum number of samples misclassified at a best V_k point is called the

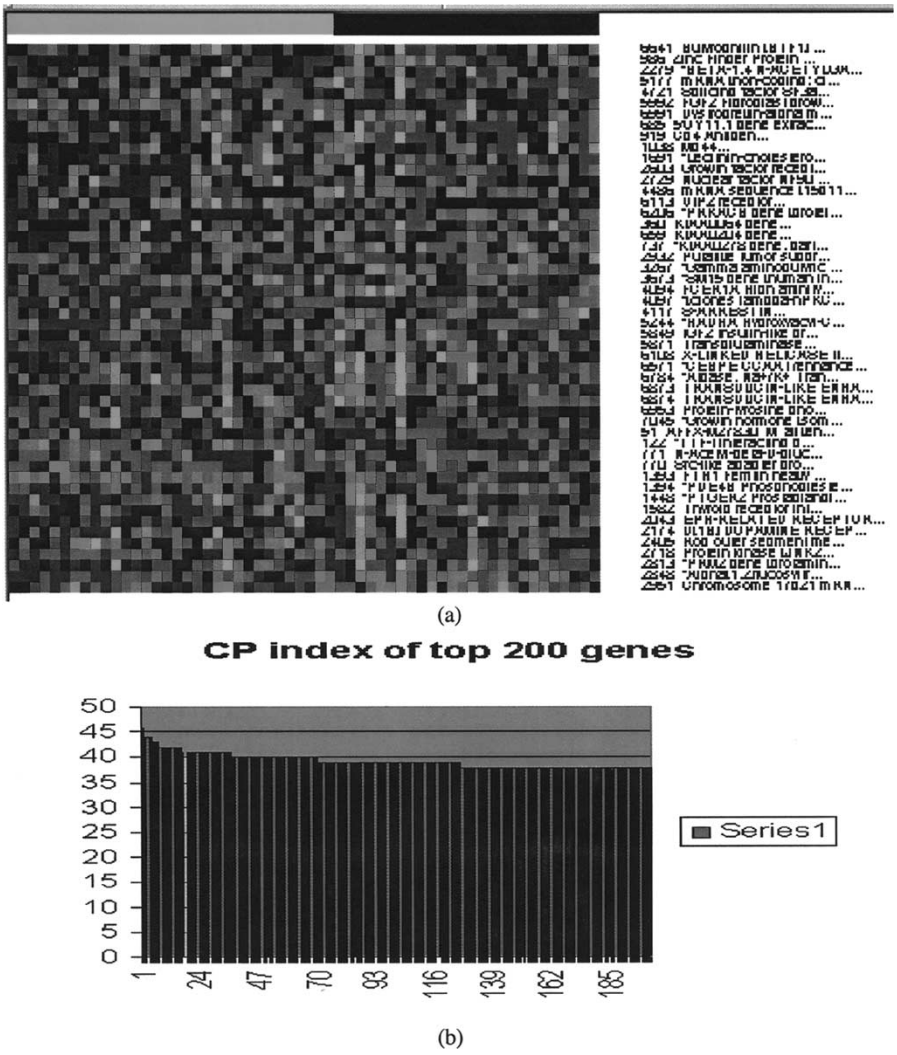

Fig. 2. (a) Genes of top 50 CPI ranks. (b) CPI values of top 200 genes.

DP index (DPI). The process of DPI computation is described as follows.

For each gene g_j its expression values over all test samples are on the row j of the data set

1) Sort the expression values of row j in ascending order;
2) Begin Loop:

k = 1;
V[k] = the kth sample-value in the sorted list (from left to right);
A[k] = number of class ω_1 samples at the left of V[k] + number of class ω_2 samples at the right of the V[k];
B[k] = number of class ω_2 samples at the left of V[k] + number of class ω_1 samples at the right of the V[k];
K++;

End of Loop;

3) DPI [j]=the smallest value from A[k] and B[k];

End of procedure.

The algorithm works in the following fashion: 1) the values in A[k] and B[k] are the possible number of misclassifications while applying V[k] as a threshold value; 2) the value DPI[j] is the minimum number of misclassifications for gene g_j; and 3) the genes are scored with the best possible nonmisclassifications (minimum misclassification) when the algorithm terminates. Note that the genes are not uniquely ranked in DPI because multiple genes can have the same DPI value, that is, scored the same. In this case, we give the same ranking value to these genes. That is, these genes will be ranked the same. Fig. 3(a) shows a graphical display of the expression profiles for the top 50 genes with high DPI values in the DLBCL dataset. The DPI values of the top 200 genes are shown in Fig. 4(b).

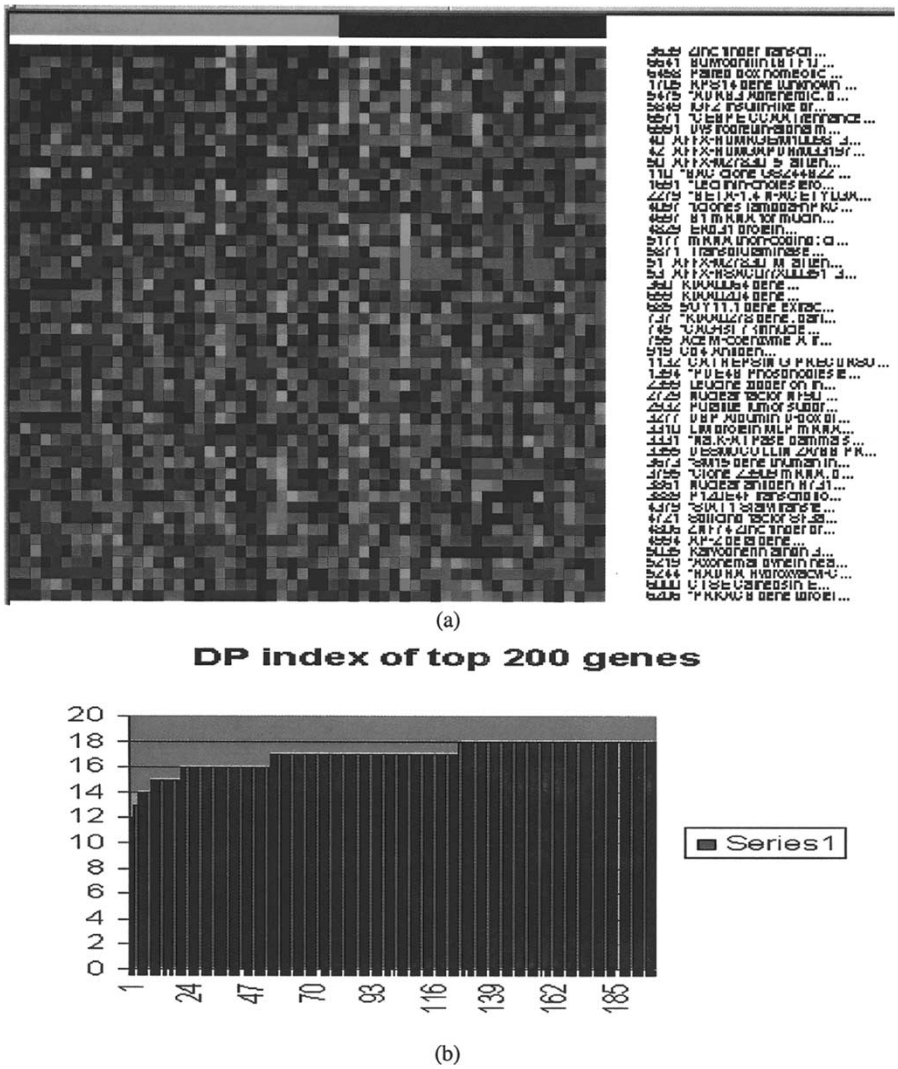

Fig. 3. (a) Genes of top 50 DPI ranks. (b) DPI values of top 200 genes.

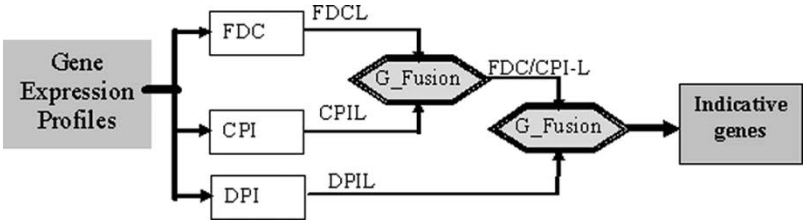

Fig. 4. Algorithmic fusion of the gene expression measurements.

D. Discussions of the Multiple Measurements

The three genes evaluation approaches, namely 1) the FDC measurement, 2) the CPI, and 3) the DPI, address different aspects of the gene evaluation criteria discussed at the beginning of this section. A comparison of the main characteristics of these approaches is shown in Table I.

Our experiment results show that the FDC, CPI, and DPI measurements do not correlate to each other on the given data set. On the other hand, genes with consistently higher ranks in all FDC, CPI, and DPI measurements do show a distinctive pattern for separating the two outcome classes. This means that by certain ways of combining the measurements, it is possible to select a set of genes that possess an overall better performance than the sets extracted from each individual measurement. Table II lists the FDC, CPI, and DPI values of four typical genes in our experiment. A column at right also indicates whether the gene is in the final selection of predictive set after applying the algorithmic fusion process to the measurements. However, the combination cannot be done by a simple application of majority voting scheme. An additive property does not hold among these measurements either. A more subtle procedural approach, thus, is needed to take count of the different measurement types and scales of the above and to come up with an integrated list of resulting genes.

TABLE I
NATURE OF THE MEASUREMENT OF THE THREE ALGORITHMS

	Statistic	Non-statistic	Continuous	Discrete	Model-based	Modeless
FDC	X		X			X
CPI	X			X	X	
DPI		X		X		X

TABLE II
FDC, CPI, AND DPI MEASUREMENTS OF A SELECTIVE SET OF GENES

Gene #	Accession No	FDC value	FDC rank	CPI value	CPI rank	DPI value	DPI rank	Selected by fusion?
3639	U70663	0.2197	1300	40	36	12	1	No (Low FDC value)
3861	U83908	0.4821	43	40	36	15	9	yes
5064	Z15114	0.5870	3	38	125	18	126	No (High DPI)
6206	M18255	0.6222	1	42	8	16	20	yes

III. ALGORITHMIC FUSION

In this section, we describe a fusion algorithm applied to the above three principal measurements for the extraction of genes that are indicative to the DLBCL clinical outcomes. Methods of combining multiple measurements and classifiers have been studied and applied successfully in solving a number of pattern recognition problems such as the handwritten character recognitions [14], [19]. Most of them dealt with classifiers that result in similar measurements or the measurements are valued in the similar nature of representations. The main feature of the fusion algorithm applied in this research is that it addresses the problem of combining multiple measurements that are in different scales and conjugative spaces of numeric values. Note that it is very likely that each gene is valued differently in the FDC measurement because of the continuity nature of the measurement. The genes, thus, can be individually ranked in the FDC measurement. However, a number of genes in the CPI and DPI measurements would be valued the same, therefore, they cannot be individually ranked according to these measurements. For example, the genes U70663 and U83908 (Table II) both have a CPI value 40, so they are both ranked at a CPI level of 36. The fusion algorithm, thus, is necessary to have a way of taking care of these cases and integrating the data that bear different ranking schemes.

A fusion algorithm called G_FUSION, which stands for algorithmic fusion of gene expression profiling measurements, is developed and applied in our research. The G-FUSION employs a parallel crossover integration technique conducted on the results of the two parametric measurement, FDC and CPI, and the one nonparametric measurement, DPI. The linkage of procedures is illustrated in the diagram shown in Fig. 4.

Let FDCL, CPIL, and DPIL denote three gene lists formed from the FDC, CPI, and DPI measurements, respectively. Note that genes in FDCL are ranked individually, while in the CPIL and DPIL, there could be multiple genes occupying the same ranking node/slot of the list. That is, each node of the CPIL or DPIL represents a set of genes at the same CPI or DPI level. Thus, the number of nodes in the FDCL is equal to the number of genes, while the number of nodes in the CPIL or DPIL is equal to the number of different CPI or DPI measurement levels. We also give a numeric value $\mathrm{Counter}(k)$ to each node k of CPIL and DPIL to record the number of genes that have the ranks fall ahead of and up to the level k of the CPI or DPI measurement

$$\mathrm{Counter}(k) = \sum_{i=0}^{k} \text{number of genes in node } i \text{ of CPIL or DPIL, respectively.}$$

The G_FUSION is a best-first rank-preserving data fusion algorithm. It means that during the fusion process, the genes with the best evaluation values in both measurements are always extracted first, and the genes in the resulting list are still ranked according to their ranks in the orders of the original lists. The fusion process runs in two steps: First, the FDCL and the CPIL are fused to generate a list FDC/CPI-L. In the second step, the FDC/CPI-L is fused with the DPIL, which results a final list of genes where the more indicative genes to the clinical outcomes are placed near the top of the list.

The fusion algorithm is presented as follows.

G_FUSION procedure: Lists Crossover (take the crossover of FDCL and CPIL as example)

Input:
1) List FDCL where genes are sorted according to the FDC measurements in descending order.
2) List CPIL where genes are sorted according to the CPI in descending order.

Output: A new list FDC/CPI-L where genes are sorted by FDC values after crossover operations on the lists FDCL and CPIL.

Computation steps:
1) Let ***d*** be a prespecified constant indicating the number of genes to be selected.
2) Initialize the list FDC/CPI-L to empty.
3) Let ***k*** be a variable initialized to the top-most level of CPI;

TABLE III
GENES EXTRACTED FROM THE ALGORITHMIC FUSION (SORTED IN CPI)

	Gene	Accession No	GeneName	FDC rank	CPI value	DPI value
1	6641	U90543	Butyrophilin (BTF1) mRNA	147	46	12
2	5177	Z49995	mRNA (non-coding; clone h2A)	136	44	15
3	6991	U46744	Dystrobrevin-alpha mRNA	29	43	14
4	4721	X85237	Splicing factor SF3a120	145	43	16
5	6206	M18255	PRKACB gene (protein kinase C-beta-2)	1	42	16
6	6873	M99435	TRANSDUCIN-LIKE ENHANCER PROTEIN 1	9	41	16
7	3673	U73167	SM15 gene (human interferon-related protein SM15	41	41	16
8	4097	X07109	(clones lambda-hPKC-beta[15,802]) (PRKCB1)	55	41	15
9	5849	M17863	IGF2 Insulin-like growth factor 2 (somatomedin A)	86	41	14
10	7045	J00148	Growth hormone (somatotropin, GH1) gene	98	41	16
11	5244	D16480	HADHA Hydroxyacyl-Coenzyme	265	41	16
12	1394	L20971	PDE4B Phosphodiesterase 4B, cAMP-specific	2	40	16
13	3861	U83908	Nuclear antigen H731 mRNA	43	40	16

4) While the number of genes in FDC/CPI-L is less than $\boldsymbol{d}$, do
   a) For each gene g_i such that g_i has the CPI value less than or equal to $\boldsymbol{k}$ in CPIL, AND g_i is NOT in the FDC/CPI-L list yet; if g_i is present in the top $\mathrm{Counter}(\boldsymbol{k})$ number of genes in FDCL, place g_i in the list FDC/CPI-L.
   b) Increase the $\boldsymbol{k}$ to the next level of the CPIL.
5) Sorting the genes in FDC/CPI-L according to their FDC values.

Note that the number of genes finally selected in the FDC/CPI-L may be greater than $\boldsymbol{d}$, because there are multiple genes at the same $\boldsymbol{k}$ rank in the CPIL. Once the $\boldsymbol{k}$ rank is reached, all genes that have the CPI value equal to $\boldsymbol{k}$ are selected with respect to their position in FDCL, which is done in step 4(a).

The same procedure is applied for a crossover fusion of the lists FDC/CPI-L and DPIL. To ensure that the second fusion step extracts sufficient number of genes ($\geq \boldsymbol{d}$), the first fusion step should have a relatively large $\boldsymbol{d}$ value. On the other hand, of course, we can always come back to the first fusion step to get more genes if the second fusion step does not generate the desired number of genes. This adaptation process is not difficult to implement on the base of the G-FUSION algorithm.

IV. EXPERIMENTAL RESULTS

The DLBL data sets used in our experiment contains 58 patient cases with 7139 genes evaluated for each case (www.genome.wi.mit.edu/MPR/lymphoma). To create an outcome predictor that integrates a number of highly predictive genes, we focused on evaluating individual genes to identify those genes that are highly predictive. The results we obtained after applying the algorithmic fusion to the total genes and cases are listed in Table III. By a selection of parameters of $d = 13$ in the second fusion procedure, a set of 13 genes with the measurements $\mathrm{FDC} - \mathrm{value} < 0.374\,364$, $\mathrm{CPI} > 39$, and $\mathrm{DPI} < 18$ were extracted.

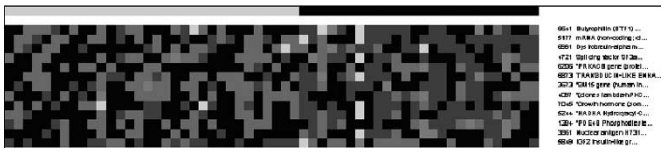

Fig. 5. Plot of gene expression values of the 13 genes extracted by the algorithmic fusion.

Using the algorithmic fusion approach, four genes in our selected set overlap with the result in Shipp's paper [17]. They are

6206 PRKACB gene (protein kinase C-beta-2);
6873 TRANSDUCIN-LIKE ENHANCER PROTEIN 1;
1394 PDE4B Phosphodiesterase 4B, cAMP-specific;
3861 Nuclear antigen H731 mRNA.

A graphical display of the gene expression values of the 13 genes in our selection is shown in Fig. 5. The mean and variance values of the genes in the predictor set based on the expression values of the original dataset are shown in Table IV.

To verify the results, we applied a simple linear discriminate function (without considering the distribution variance and assuming equal probability for both survival and fatal cases) to the DLBCL gene expression profiling dataset. The test yields a total of 46 (27 + 19) cases correctly identified. That is, one additional case was correctly assigned in our experiment compared with Shipp's 13 genes reported [17] (see Table V). The linear classifier (predictor) takes a form of the following:

$$\begin{aligned} G(\underline{x}) &= g_1(x) - g_2(x) = \|\underline{x} - \underline{\mu}_1\|^2 - \|\underline{x} - \underline{\mu}_2\|^2 \\ &= \left(\underline{x}^T\underline{x} - 2\underline{\mu}_1^T\underline{x} + \underline{\mu}_1^T\underline{\mu}_1^T\right) - \left(\underline{x}^T\underline{x} - 2\underline{\mu}_2^T\underline{x} + \underline{\mu}_2^T\underline{\mu}_2^T\right) \\ &= -2\left(\underline{\mu}_1^T - \underline{\mu}_2^T\right)\underline{x} + \left(\underline{\mu}_1^T\underline{\mu}_1^T - \underline{\mu}_2^T\underline{\mu}_2^T\right). \end{aligned}$$

where $\underline{x} = [x_1, x_2, \ldots, x_{15}]$ is a normalized vector that contains the 15 genes of the extracted gene set. The $\underline{\mu}_1$ and $\underline{\mu}_2$ are the mean vectors of $\underline{x}$ with respect to the two classes (survival–fatal), respectively.

TABLE IV
MEAN AND VARIANCE VALUES OF THE 13 GENES EXTRACTED FROM THE ALGORITHMIC FUSION

	Gene #	Accession No	Over all cases			
			Survival (0)		Non-survival (1)	
			Mean (μ_1)	Variance (σ_1)	Mean (μ_2)	Variance (σ_2)
1	6641	U90543	-215.906	105.323	-134.808	171.346
2	5177	Z49995	-140	64.9822	-89.1154	102.229
3	6991	U46744	-159.875	60.253	-100.731	86.9481
4	4721	X85237	-43.3438	168.339	65.8846	209.477
5	6206	M18255	236.406	119.342	497.231	460.373
6	6873	M99435	-111.313	91.4455	-12.6923	134.74
7	3673	U73167	93.8125	271.468	282.654	228.143
8	4097	X07109	1560.88	1312.02	2665.5	2156.47
9	5849	M17863	258.313	69.6937	202.846	98.9365
10	7045	J00148	-293.688	119.166	-213.885	151.83
11	5244	D16480	12.75	635.134	401.5	826.876
12	1394	L20971	250.188	190.535	535.808	406.379
13	3861	U83908	-6.75	71.4383	78.3077	172.685

TABLE V
CLASSIFICATION TESTS ON THE EXTRACTED GENE SET

		Linear classifier			Quadratic classifier		
		survival	fatal	total	survival	fatal	total
1	With use of our 13 genes	27	19	**46**	30	18	**48**
2	With use of Shipp's 13 genes	26	19	45	26	19	45

Applying a quadratic discriminate function (with consideration of the distribution variance parameters and assuming equal probability for both survival and fatal cases) defined on the extracted genes to the DLBCL gene expression profiling dataset yields a total of 48 (30 + 18) cases correctly identified. That is, three additional cases were correctly assigned compared with Shipp's report. The quadratic classifier (predictor) takes a form of the following:

$$
\begin{aligned}
G(\underline{x}) &= g_1(x) - g_2(x) \\
&= \left\{ -\frac{1}{2}\ln\left(|\underline{K}_1|\right) - \frac{1}{2}[\underline{x}-\underline{\mu}_1]^T \underline{K}_1^{-1}[\underline{x}-\underline{\mu}_1] \right\} \\
&\quad - \left\{ -\frac{1}{2}\ln\left(|\underline{K}_2|\right) - \frac{1}{2}[\underline{x}-\underline{\mu}_2]^T \underline{K}_2^{-1}[\underline{x}-\underline{\mu}_2] \right\} \\
&= \left\{ \left[-\frac{1}{2}\left(\underline{x}^T \underline{K}_1^{-1}\underline{x} - \underline{x}^T \underline{K}_1^{-1}\underline{\mu}_1 \right.\right.\right. \\
&\qquad \left.\left.\left. -\underline{\mu}_1^T \underline{K}_1^{-1}\underline{x} + \underline{\mu}_1^T \underline{K}_1^{-1}\underline{\mu}_1 \right) \right] \right. \\
&\quad \left. - \left[-\frac{1}{2}\left(\underline{x}^T \underline{K}_2^{-1}\underline{x} - \underline{x}^T \underline{K}_2^{-1}\underline{\mu}_2 \right.\right.\right. \\
&\qquad \left.\left.\left. -\underline{\mu}_2^T \underline{K}_2^{-1}\underline{x} + \underline{\mu}_2^T \underline{K}_2^{-1}\underline{\mu}_2 \right) \right] \right\} \\
&\quad + \left\{ \left[-\frac{1}{2}\ln\left(|\underline{K}_1|\right) \right] - \left[-\frac{1}{2}\ln\left(|\underline{K}_2|\right) \right] \right\} \\
&= -\frac{1}{2}\underline{x}^T \left(\underline{K}_1^{-1} - \underline{K}_2^{-1} \right) \underline{x} + \underline{x}^T \left(\underline{K}_1^{-1}\underline{\mu}_1 - \underline{K}_2^{-1}\underline{\mu}_2 \right) \\
&\quad - \left[\underline{\mu}_1^T \underline{K}_1^{-1}\underline{\mu}_1 - \underline{\mu}_2^T \underline{K}_2^{-1}\underline{\mu}_2 \right) + \frac{1}{2}\ln\left(\frac{|\underline{K}_1|}{|\underline{K}_2|} \right) \Bigg].
\end{aligned}
$$

which can be written as

$$G(\underline{x}) = \underline{x}^T \underline{\Psi} \underline{x} + \underline{W}^T \underline{x} + W_0$$

where

$\underline{\Psi}_i \quad -(1/2)(\underline{K}_1^{-1} - \underline{K}_2^{-1})$,

$\underline{W}^T \quad (\underline{K}_1^{-1}\underline{\mu}_1 - \underline{K}_2^{-1}\underline{\mu}_2)^T$, and

$W_0 \quad -[\underline{\mu}_1^T \underline{K}_1^{-1}\underline{\mu}_1 - \underline{\mu}_2^T K_2^{-1}\underline{\mu}_2) + (1/2)\ln(|\underline{K}_1|/|\underline{K}_2|)]$.

The $\underline{K}_1$ and $\underline{K}_2$ are the covariance matrices of $\underline{x}$ with respect to the two classes (survival–fatal), respectively.

Note that it is not always true that the gene expression values are independent to each other. That is, there are some genes such that their expression values with respect to certain clinical aspect, such as the survival–fatal outcome, are correlated. The genes can tend to be coactivated in response to a given stimulus and, therefore, cannot guarantee to be expressed independently of one another. The implication of this correlation means that some of the genes in the set extracted in our experimentation may be correlated. Measurement theory suggests that the relationships among items are logically connected to the relationships of items to the latent variable. Therefore, additional analysis and gene pattern classification processes performed on different clinical aspects of the genes are needed to detect and identify the correlations (or irrelevance) of these genes. However, this does not exclude us to use the extracted set of genes as the outcome predictors. There are other technical aspects concerning the types of measurements and the uniqueness of the measurements, in relation to the general measurement theory that address the dependency problem and need to be further studied.

V. CONCLUSION

We have presented a multifaceted measurement integration approach for identification of genes predictive to patient's survival outcomes in the analysis of a DLBCL expression profiling dataset. A main feature of the approach is that it combines the measurements in both parametric and nonparametric domains, as well as continuous and discrete natures. The algorithmic fusion attempts to attenuate the effects of measurement uncertainties coherent in the dataset by proper association of the multifaceted measurements. However, we must point out that our results were purely based on numeric evaluations of the gene expression profiles. No considerations of the biological nature of the genes are taken; that is, the functionalities of

the genes are not considered in the selection process. Moreover, these measurements provide the relative importance, in terms of the discrepancies of the expression values between cases of different classes (survival–fatal), of the genes with respect to the others. Since we limited the selection to 13 genes, it is inevitable that some important genes, many of which are possibly biologically meaningful, may be missed in our results. We are sure that the gene list can be further expanded with an adjustment of the selection parameters, such as the value d of the fusion algorithm. We also believe that our method can be applied to many other data sets analyses tasks, including the gene expression profiles for other clinical diagnosis assistances.

REFERENCES

[1] A. Alizadeh and M. Eisen *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 4051, pp. 503–511, 2000.

[2] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, 1979.

[3] A. Ben-Dor and A. Yakhini, "Clustering gene expression patterns," *J. Comput. Biology*, vol. 6, pp. 281–297, 1999.

[4] M. Brown *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," in *Proc. National Academy Sciences*, vol. 97, 2000, pp. 262–267.

[5] R. F. DeVellis, *Scale Development: Theory and Applications*, 2nd ed. Newbury Park, CA: Sage, 2003.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugenics*, pt. II, vol. 7, pp. 179–188, 1936.

[8] J. H. Friedman, "Exploratory projection pursuit," *J. Amer. Statistical Assoc.*, vol. 82, no. 397, pp. 249–266, 1987.

[9] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.*, vol. C-23, pp. 881–890, 1974.

[10] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[11] ——, "Projection pursuit," *Ann. Statistics*, vol. 13, no. 2, pp. 435–475, 1985.

[12] G. P. Nason, "Design and choice of projection indices," Ph.D. thesis, Univ. Bath, 1992.

[13] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Netw.*, vol. 5, pp. 927–935, 1992.

[14] B. Plessis, A. Sicsu, and L. Heutte *et al.*, "A multi-classifier combination strategy for the recognition of handwritten cursive words," in *Proc. Int. Conf. Document Analysis and Recognition*, 1993, pp. 221–226.

[15] C. Posse, "Projection pursuit discriminant analysis for two groups," *Commun. Statist. Theory Methods*, vol. 21, no. 1, pp. 1–19, 1992.

[16] A. Rosenwald, G. Wright, and W. Chan *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma," *New Eng. J. Med.*, vol. 346, no. 25, pp. 1937–1947, 2000.

[17] M. A. Shipp, K. N. Ross, and P. Tamayo *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[18] P. Suppes and J. L. Zinnes, "Basic measurement theory," in *Mathematical Psychology*, R. D. Luce, R. R. Bush, and E. Galanter, Eds. New York: Wiley, 1963, pp. 4–76.

[19] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418–435, May/June 1992.

[20] J. Yang, J. Yang, and D. Zhang, "What's wrong with Fisher criterion?," *Pattern Recognit.*, vol. 35, no. 11, pp. 2665–2668, 2002.

Qiuming Zhu (SM'97) received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1986.

He is a Professor of Computer Science at the University of Nebraska at Omaha, Omaha. He was a postdoctoral Research Fellow in the Center for Computer Aids for Industrial Productivity at Rutgers University, New Brunswick, NJ, and an Assistant Professor of Computer Science and Engineering at Oakland University, from 1986 to 1990. His research interests are in digital image processing and computer vision, pattern recognition, neural networks, data fusion and data mining, and artificial intelligence applications in science and engineering.

Hongmei Cui received the B.S. degree in East China Normal University, Shanghai, China, and her M.S. degree in computer science from the University of Nebraska at Omaha, Omaha.

Her research interests are in database management, data mining, knowledge-based systems, bioinformatics, and software development in biomedicine. She is currently working at Omaha World-Herald.

Kajia Cao received the B.E. degree in computer science and engineering from Southeast University, Nanjing, China, and the M.S. degree in computer science from the University of Nebraska at Omaha, Omaha. She is currently pursuing the Ph.D. degree at the University of Nebraska at Omaha.

Her research interests are in the areas of data mining and knowledge discovery, bioinformatics, and advanced algorithms of computing in biomedicine.

Wing (John) C. Chan received his pathology training at the University of Chicago and specializes in hematopathology.

He is a Professor of Pathology at the University of Nebraska Medical Center. His current research interests focus on discovering important genetic abnormalities that determine the biologic and clinical behaviors of Hodgkin and non-Hodgkin lymphoma. An important aspect of this effort is the gene expression profiling of a large series of lymphomas through the collaborative efforts of UNMC with a number of institutions in the United States and Europe. He and his collaborator, Dr. Staudt from the NCI, are co-chairs of the coordinating committee overseeing this project which is funded in part by a grant from the National Institutes of Health.

Dr. Chan has served in a number of NIH study sections and as an Associate Editor of the American Journal of Pathology. He is currently on the editorial board of the *American Journal of Clinical Pathology* and the *American Journal of Pathology*.