Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Group classification based on high-dimensional data: application to differential scanning calorimetry plasma thermogram analysis of cervical cancer and control samples

This article was published in the following Dove Press journal:
Open Access Medical Statistics
16 April 2013
Number of times this article has been viewed

Shesh N Rai[1,2]
Jianmin Pan[1]
Alex Cambon[2]
Jonathan B Chaires[3–5]
Nichola C Garbett[3,4]

[1]Biostatistics Shared Facility, James Graham Brown Cancer Center, University of Louisville, [2]Department of Bioinformatics and Biostatistics, University of Louisville, [3]Biophysical Core Facility, James Graham Brown Cancer Center, University of Louisville, [4]Department of Medicine, University of Louisville, [5]Department of Biochemistry and Molecular Biology, University of Louisville, Louisville, KY, USA

Abstract: Differential scanning calorimetry has been applied to identify protein denaturation patterns, or thermograms, in blood plasma samples that are indicative of health status. Data sets generated by differential scanning calorimetry are high dimensional, and it is complex to analyze and classify thermogram patterns. The I-RELIEF method is commonly used for group classification from high-dimensional data sets, such as gene expression data. We report the development and validation of a new method of data reduction and modeling of high-dimensional data sets. The performance of our method was demonstrated through its application to the analysis of differential scanning calorimetry plasma thermogram data. Our method was found to provide superior classification performance compared with the I-RELIEF method.

Keywords: plasma thermogram, differential scanning calorimetry, group classification

Introduction

The goal of classification is to predict class outcomes. Existing data of known classes are used to build a model. This model is then applied to predict class outcomes on data where the class outcomes are blinded or not yet known. Some of the first developed and still commonly used classification methods are linear discriminant analysis (LDA),[1] logistic classification,[2] and nearest neighbors.[3] LDA is closely related to linear regression and analysis of variance (ANOVA). Logistic classification is a generalized linear model with fewer assumptions than LDA. One of the challenges of classification involves what is known as high-dimensional data, that is, data in which the number of features is much greater than the sample size. It is well known that, for example, genomic or proteomic data often include many features but that only a few of these may be relevant. In this situation, use of effective feature selection or feature reduction is needed during the classification process.

The three classification methods mentioned briefly above all require separate feature selection/reduction methods in the presence of high-dimensional data. Recently developed methods which have built-in procedures to address this issue include Random Forests, Support Vector Machines and Boosting.[4–7] Boosting and Random Forests are also ensemble methods, since they use a combination of many models to improve prediction and reduce classification error. The method proposed in this work is an ensemble method, as are some extensions of the feature selection and classification algorithm RELIEF, proposed by Kira and Rendell.[2] Dietterich[8] considered RELIEF-F to be one of the most successful preprocessing algorithms because of its

Correspondence: Shesh N Rai
Clinical and Translational Research Building, Room 211, 505 South Hancock Street, Louisville, KY 40202, USA
Tel +1 502 852 4030
Fax +1 502 852 7979
Email shesh.rai@louisville.edu

http://dx.doi.org/10.2147/OAMS.S40069

random sampling mechanism and because of the way in which it selects and weights features. However, the main drawback of RELIEF is that it makes an implicit assumption that the nearest neighbors found in the original feature space are the ones in the weighted space, which is highly unlikely in practical application and lacks a mechanism to eliminate outlier data.[3] Hence, Sun and Li[3] proposed an analytic solution, I-RELIEF, to resolve these two issues; also, I-RELIEF converges to a unique solution regardless of initial starting points, under certain conditions. Further, Steinberg et al[9] developed a method that allowed for optimal allocation of the sample data sets between case and control cohorts as well as computing sample size, when the goal of the study is to prove that the test procedure exceeds prestated bounds for positive and negative predictive values (PPV and NPV). In this paper, we propose a new ensemble classification method, which uses an extended linear model approach together with random sampling without replacement.

Classification error is an overall measure that can be used to compare different classification methods.[4] However, when predicting classes, such as disease status, there may be different consequences associated with incorrectly classifying an individual. For this reason, classification methods are often characterized by more specific measures, such as sensitivity (Sens), specificity, (Spec), PPV, and NPV. We use these characteristics to compare our new classification method with the I-RELIEF method.

This paper is organized in the following way: The next sections describe first, the motivating example and then, the assessment of the performance of the classification method. The existing method, I-RELIEF, is then introduced, followed by the description of our proposed method. We then detail the application of both I-RELIEF and our new method to the analysis of differential scanning calorimetry (DSC) thermogram data and provide discussion of the results. The overall conclusions and future direction is given in the last section.

The motivating example

We have recently applied DSC analysis of human blood plasma as a method to detect and monitor disease-related changes in the plasma proteome.[10–13] DSC analysis in biochemistry is a thermoanalytical technique that monitors the thermal properties of biomolecular solutions as a function of temperature. The technique directly and precisely measures the heat capacity of thermal events, yielding a profile known as a thermogram, which is characteristic for a given biomolecule. DSC is extremely sensitive to the precise composition of biomolecular mixtures, with the observed signal related to the amount and interaction of component biomolecules. It is this characteristic that forms the basis of the utility of DSC as a monitor of health status, via the detection of changes in the blood plasma proteome associated with the development of human disease. Figure 1 shows a schematic of a DSC instrument and thermogram output. Two identical chambers (sample and reference) are heated at a controlled rate, under constant pressure, by the main instrument heaters. The temperature difference between these chambers (ΔT), resulting from thermal events occurring in the sample cell, is monitored and electrical power applied to feedback heaters to keep the chambers in thermal balance. The feedback power signal provides a direct measurement of the heat capacity of thermal events occurring within the DSC chamber. For the ideal case of the two-state thermal denaturation of a pure single-domain protein, the DSC thermogram is represented by a unimodal melting curve, where the midpoint temperature corresponds to the melting temperature of the protein, and the area represents the denaturation enthalpy. For multiple domain proteins or mixtures of proteins, the thermogram appears more complex, with composite features of the proteins in the test solution. Data from our laboratory has shown that the thermogram of plasma from healthy "normal" individuals reflects the sum of thermograms of the component proteins, weighted according to their abundance in plasma.[10] Plasma thermograms from patients suffering from a variety of diseases appear different in amplitude and denaturation temperature.[10,13] Preliminary data show that these differences correlate with the type and stage of disease, and we hypothesize that these are related to disease-specific changes of concentration, modification, or intermolecular interaction of components, within the plasma proteome. For some diseases, dramatic alteration of the plasma thermogram relative to that of normal individuals would make qualitative identification

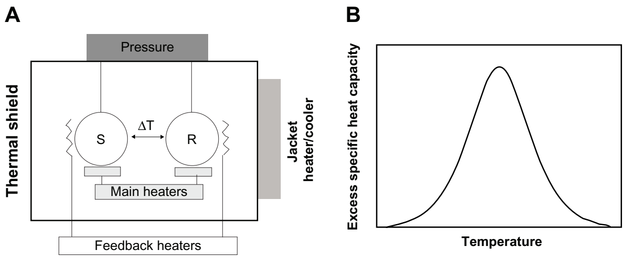

Figure 1 (**A**) Schematic of a DSC instrument (S and R refer to the sample and reference chambers of the instrument (see text); (**B**) DSC thermogram for an ideal two-state denaturation of a pure single domain protein.

Abbreviations: DSC, differential scanning calorimetry; R, reference; S, sample; ΔT, temperature change.

of disease status trivial. For other diseases, changes are less dramatic and disease classification is challenging. The motivation behind the analyses described in this article was to develop an approach for the classification of plasma thermograms according to disease status, an essential step in the development of the clinical utility of DSC thermograms. Specifically, the objective was to develop a method to classify subjects between control and case groups, based on the characteristics of thermogram data sets. The performance of the method was demonstrated through its application to two clinical groups, commercially obtained plasma samples from healthy normal individuals (the Normal group) and plasma specimens obtained from patients attending the Division of Gynecologic Oncology clinic of the James Graham Brown Cancer Center (the Cervical group).

Plasma thermogram data was obtained for the two clinical groups according to our previously published procedure.[10] Briefly, small aliquots of plasma samples (100 μL) were dialyzed for 24 hours at 4°C against a standard phosphate buffer (10 mM potassium phosphate, 150 mM sodium chloride, 15 mM sodium citrate, pH 7.5), to ensure normalization of buffer conditions. After dialysis and before use, recovered samples and final dialysis buffer were filtered. Samples were diluted 25-fold with dialysis buffer, to obtain a suitable concentration for DSC analysis. Thermograms were collected using an automated DSC instrument at 0.1°C increments over the temperature range 20°C–110°C, normalized for the total protein content of the sample and analyzed using the instrument-supplied software, to yield final thermogram data in the form of excess specific heat capacity (cal/°C-g) versus temperature (°C). For each sample, duplicate thermogram measurements were performed. Thermograms were truncated to a temperature range of 45°C–90°C which encompassed denaturation profiles of all major plasma proteins. Thus, each thermogram comprised 451 data points and, with duplicate measurements, 902 data points for each subject.

The Normal group comprised 100 commercial plasma samples purchased from Innovative Research (Novi, MI, USA) . The demographic characteristics were: 25 Caucasian males, 25 Caucasian females, 15 Hispanic males, 15 Hispanic females, ten African-American males and ten African-American females, ranging in age from 18–61 years (mean 35.8 years). Three samples were excluded from subsequent analysis as a result of sample handling or technical issues, giving 97 samples in the Normal group. Even though male subjects were not represented in our second clinical group, that of cervical disease, we decided not to exclude these data sets for two reasons: (1) we have observed little effect of gender on the thermogram profile, and (2) we wanted to incorporate the use of a general Normal control set for use in the subsequent development of our classification methods. The Cervical group consisted of plasma specimens obtained from the Division of Gynecologic Oncology Tissue Bank in the James Graham Brown Cancer Center at the University of Louisville, KY as part of a study reviewed and approved by the University of Louisville Institutional Review Board (Study 08.0108). Samples represented patients with different stages of invasive cervical cancer and different grades of precancerous cervical lesions. A total of 44 patient specimens were analyzed during a pilot study to evaluate the utility of DSC as a clinical diagnostic tool for the detection of cervical disease. Demographic characteristics were: 33 Caucasian, three Hispanic, seven African-American, and one Vietnamese, with an age range of 18–66 years (mean 35.5 years). Five patients were excluded from subsequent analysis: four patients for whom the primary site of the lesion or cancer was not confined solely to the cervix, and one patient who represented a very early stage of precancerous lesion for which there was only one sample. This gave 39 samples in the Cervical group.

Performance measurement of classification methods

The ability to correctly classify a test sample can be measured by Sens, Spec, PPV, and NPV, which are defined in Table 1.

Suppose A samples from $A + C$ and D samples from $B + D$ were correctly classified into case and control groups, respectively; then, we define Sens as $A/(A + C)$ and Spec as $D/(B + D)$. Sens and Spec represent the fraction of true case and true control classification results, respectively and are common metrics describing the performance of a diagnostic method. Similarly, PPV can be defined as $A/(A + B)$ and NPV as $D/(C + D)$. PPV (NPV) represents the chance that a sample classified into the case (control) group corresponds to an actual case (control) sample. The classification rate can also be calculated as $(A + D)/(A + B + C + D)$. Higher values for each of these statistics indicate higher quality diagnostic performance.

Table 1 Measurement of classification methods

Group	True		Total
	Case	Control	
Classified			
Case	A	B	$A + B$
Control	C	D	$C + D$
Total	$A + C$	$B + D$	$A + B + C + D$

Introduction to the existing method, I-RELIEF

Sun and Li[3] introduced I-RELIEF to identify a hybrid signature through the combination of both genetic and clinical markers of gene-expression data. They concluded that I-RELIEF performs significantly better than other methods, including the 70-gene signature, clinical markers alone, and the St Gallen consensus criterion. We outline the I-RELIEF method below.

Let $D = \{(Y_n, G_n), n = 1, \ldots, N\}$ denote a training dataset, where Y_n is the *n*th data sample (a vector valued) and $G_n \in \{1,0\}$ is the group indicator, ie, case or control. The *i*th component of Y_n is the *i*th measurement in the *n*th sample. A margin for the sample Y_n is defined as $\rho_n = d(Y_n - NM(Y_n)) - d(Y_n - NH(Y_n))$, where NM($Y_n$) and NH($Y_n$) are the nearest miss and nearest hit of Y_n, which can be regarded as two functions that, given an input Y_n, return the nearest neighbors of Y_n from the opposite and same classes, respectively; $d(\cdot)$ is a distance function defined as $d(Y) = \sum_{i=1}^{K} | y_i |$, where K is a dimension of Y. Note that $\rho_n > 0$ only if Y_n is correctly classified by a one-nearest-neighbor classifier. Then the averaged margin in a weighted feature space is maximized as follows:

$$\max_W \sum_{n=1}^{N} \rho_n(W)$$
$$= \max_W \sum_{n=1}^{N} \sum_{i=1}^{K} w_i \left\{ | Y_n^{(i)} - NM^{(i)}(Y_n) | - | Y_n^{(i)} - NH^{(i)}(Y_n) | \right\}$$

with $\|W\|_2^2 = 1$ and $W \geq 0$, where $\rho_n(W)$ is the margin of Y_n computed with respect to W, where W is a weight vector, which needs to be estimated. It has been proven that the optimization scheme in the above equation can be solved with a closed-form solution and is equivalent to the well-known RELIEF algorithm.[2,3] One major drawback of RELIEF, however, is that the nearest neighbors are defined in the original feature space, which is highly unlikely to be the ones in the weighted space. In the presence of many irrelevant features, which is the case in microarray data analysis, the performance of RELIEF can degrade significantly. Hence Sun and Li[3] proposed I-RELIEF, which provides an analytic solution to mitigate the problem of RELIEF.

Define two sets $M_n = \{i: 1 \leq i \leq N, G_i \neq G_n\}$ and $H_n = \{i: 1 \leq i \leq N, G_i = G_n, \mathrm{i} \neq n\}$ associated with each sample Y_n. Suppose the nearest hit and miss are known for each sample and the indices of which are recorded in the set $S_n = (s_{n1}, s_{n2})$, where $s_{n1} \in M_n$ and $s_{n2} \in H_n$. Then the objective function to be optimized can be formulated as

$$C(W) = \sum_{n=1}^{N} \left(\left\| Y_n - Y_{S_{n1}} \right\|_W - \left\| Y_n - Y_{S_{n2}} \right\|_W \right),$$

where $\|Y\|_W = \sum_{i=1}^{K} w_i \, | y_i |$ and the equation can be easily optimized using RELIEF. However, we do not know the set $S = \{S_n, n = 1, \ldots, N\}$. By following the principle of the Expectation Maximization algorithm, the elements of $S = \{S_n\}$ are regarded as hidden random variables, and we can derive the probability distributions of the unobserved data. First, an estimate is made of the weight vector W. The probability of the *i*th data point being the nearest miss of Y_n if $i \in M_n$, or being the nearest hit of Y_n if $i \in H_n$, can then be defined as

$$P_m(i \mid Y_n, W) = \frac{f\left(\left\| Y_n - Y_i \right\|_W\right)}{\sum_{j \in M_n} f\left(\left\| Y_n - Y_j \right\|_W\right)}$$

and

$$P_h(i \mid Y_n, W) = \frac{f\left(\left\| Y_n - Y_i \right\|_W\right)}{\sum_{j \in H_n} f\left(\left\| Y_n - Y_j \right\|_W\right)},$$

respectively, where $f(\cdot)$ is a kernel function. One commonly used kernel function is $f(d) = \exp(-d/\sigma)$, where σ is a user-defined parameter and σ = 2 is commonly used based on empirical experience. For notational brevity, define $\alpha_{i,n} = P_m(i \mid Y_n, W^{(t)})$, $\beta_{i,n} = P_h(i \mid Y_n, W^{(t)})$, $\Omega = \{W: \|W\|_W = 1, W \geq 0\}$, $M_{n,i} = | Y_n - Y_i |$ if $i \in M_n$ and $H_{n,i} = | Y_n - Y_i |$ if $i \in H_n$. Thus, I-RELIEF can be summarized as follows:

Step 1: After the *i*th iteration, the Q function is calculated as:

$$\begin{aligned} Q(W \mid W^{(t)}) \square \ E_S C(W) &= \sum_{n=1}^{N} \Bigg(\sum_{i \in M_n} \alpha_{i,n} \left\| Y_n - Y_i \right\|_W \\ &\quad - \sum_{i \in H_n} \beta_{i,n} \left\| Y_n - Y_i \right\|_W \Bigg) \\ &= \sum_{n=1}^{N} \Bigg(\sum_j w_j \sum_{i \in M_n} \alpha_{i,n} M_{n,i}^{j} - \sum_j w_j \sum_{i \in H_n} \beta_{i,n} H_{n,j}^{j} \Bigg) \\ &= W^T \sum_{n=1}^{N} (\overline{M}_n - \overline{H}_n) \square \ W^T V \end{aligned}$$

where $\overline{M}_n = \sum_{i \in M_n} \alpha_{i,n} M_{n,i}$ and $\overline{H}_n = \sum_{i \in H_n} \beta_{i,n} h_{n,i}$.

Step 2: Estimation of W in the $(t + 1)$th iteration is:

$$W^{(t+1)} = \arg\max_{W \in \Omega} Q(W \mid W^{(t)}) = v^+ \Big/ \left\| v^+ \right\|_2 .$$

where $v_i^+ = \max(v_i, 0)$.

Alternating iteration of Steps 1 and 2 is performed until convergence is reached, ie, $\left\| W^{(t+1)} - W^{(t)} \right\| < \theta$, where θ is a small positive number. Sun and Li[3] have proven mathematically that I-RELIEF converges to a unique solution regardless of the initial weights if the kernel function is properly selected, with convergence typically achieved within a few iterations.

I-RELIEF combines the merits of both filter and wrapper methods. Note that the objective function optimized by I-RELIEF approximates the leave-one-out accuracy of a nearest-neighbor classifier. Therefore, I-RELIEF can be regarded as a wrapper method and thereby, it naturally addresses the issues of feature correlation and the removal of redundant features. Moreover, I-RELIEF can be solved analytically and thus avoids any heuristic combinatorial search. The effectiveness of the algorithm has been demonstrated through large-scale experiments on simulated data and six microarray datasets.[3]

The proposed method

In this section, we introduce a new classification method in a few steps as follows. Suppose we have a dataset with two groups of samples: a control group including n_1 samples and a case group including n_2 samples, where $n_1 + n_2 = N$. There are K measurements at K points, such as temperature, for each sample. Let Y and X be the response variable and covariate vector, respectively, and G be the group indicator, 0 for control and 1 for case.

Step 1: Compute 95% quantile of residuals at each point using N observations from all samples of both groups.

We use all measurements from N samples to fit the regression model below:

$$Y = \alpha + \beta_0 G + \beta^T X + \varepsilon. \tag{1}$$

Based on this model, we calculate the residuals for each one of $N \times K$ observations, denoted as $R_{i,j}(i = 1, \ldots, K \text{ and } j = 1, \ldots, N)$, and then calculate the 95% quantile of each of N absolute values of residuals, $R_{i,1}, \ldots, R_{i,N}$, for each temperature point $i(i = 1, \ldots, K)$, denoted as $q_1, \ldots, q_k$.

Step 2: Fit the regression model (Equation 1) using observations from m_1 control samples and m_2 case samples.

We randomly select m_1 samples from n_1 control samples and m_2 from n_2 case samples and use $m_1 + m_2$ samples to fit the regression model (Equation 1).

Step 3: Validate the classification method using the remaining $n_1 - m_1$ control samples and $n_2 - m_2$ case samples.

Based on the model (Equation 1) fitted in Step 2 using m_1 control samples and m_2 case samples, we compute the predicted observation for $(n_1 - m_1 + n_2 - m_2)$ samples, $\hat{Y}(G = 0)$ and $\hat{Y}(G = 1)$, and their prediction residuals for both $G = 0$ or $G = 1$, no matter which group the sample was from. Thus, we have $2K$ residuals for each sample data point, r_i (G = 0) for $G = 0$ and r_i (G = 1) for $G = 1$ ($i = 1, \ldots, K$). Each K residuals were compared to the quantiles derived in Step 1 for both $G = 0$ and $G = 1$, and P-values were calculated according to Equation 2 below:

$$P(G = j) = \frac{\# of \{ | r_i(G = j) | > q_i,\ i = 1, \ldots, K \}}{K} \quad \text{for } j = 0, 1. \tag{2}$$

Step 4: Compute the Sens, Spec, PPV, NPV, and the classification rate.

Step 5: Repeat steps 2 and 3 a total of B times (B = 5000 in the following application) and compute the average $P(G = 0)$ and $P(G = 1)$ values to yield $\overline{P}(G = 0)$ and $\overline{P}(G = 1)$. If $\overline{P}(G = 0) \geq \overline{P}(G = 1)$, we classify the sample into the case group; otherwise, it is classified into the control group. Note that other covariates (such as demographic, etc) can also be included in addition to thermogram data, for model building and prediction.

Application of the proposed method to the analysis of plasma thermogram data

Evaluation and data reduction

For initial comparison and analysis of thermograms, each data set collected at 0.1°C intervals over the temperature range 20°C–110°C was subsequently truncated to a temperature range of 45°C–90°C that spanned the region of interest encompassing thermal denaturation of the major plasma proteins. Figure 2 shows composite plots for all thermograms analyzed during this study, with duplicate scans for 97 Normal and 39 Cervical samples. Each thermogram comprised 451 data points and, with duplicate measurements, 902 data points for each sample. The entire data set of 97 Normal samples and 39 Cervical samples was extremely data rich, with a total of 122,672 observations from 136 samples. For development of the classification model, it was necessary to significantly reduce the complexity of this high-dimensional data set. Closer examination revealed negative or low-heat capacity values below 50°C and above 75°C, and these temperature ranges were excluded from analysis. Also, significant thermogram variation was observed within and between each group.

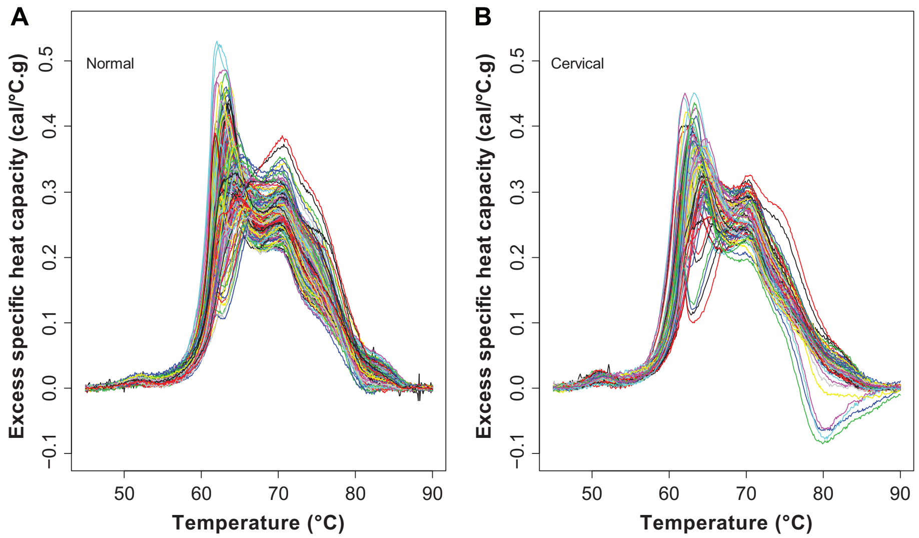

Figure 2 Composite plots of duplicate scans for (**A**) 97 Normal samples; and (**B**) 39 Cervical samples, analyzed during this study.

An estimate of the reproducibility of individual DSC measurements was provided by a comparison of duplicate sample measurements and was found to be highly reproducible compared with the biological variability of the samples. For each sample, heat capacity values were averaged for duplicate thermograms and also within each 1°C increment, ie, from 50.0°C to 50.9°C, 51.0°C to 51.9°C, and so on, to yield average heat capacity values from 20 measurements at each temperature. The result was a total of 26 observations for each sample at 1°C intervals over the temperature range of 50.0°C to 75.9°C, that is, $K = 26$, $n_1 = 39$ and $n_2 = 97$.

Figure 3 shows a composite scatter plot for Normal and Cervical samples. It can be seen that the heat capacity distributions were asymmetric and also very similar between the two groups of samples, which indicated that it was difficult to classify the two groups using general classification procedures. Because of the observed asymmetry, we transformed the original measurements and performed the Shapiro–Wilk normality test at each temperature point. Since all values of heat capacity were between 0 and 0.5, we considered five transformations denoted as H as follows: $H_1 \hat{=} \log(H)$, $H_2 \hat{=} \log it(H/0.5)$, $H_3 \hat{=} \log it(2H)$, $H_4 \hat{=} e^H/(1+e^H)$, and $H_5 \hat{=} e^{2H}/(1+e^{2H})$, where $\log(\cdot)$ is the natural logarithm. The P-values for the normality test are presented in Table 2, for the original data and each of the five transformations. The P-values values show that the logarithm transformation (H_1) performed the best for the normality test, and this transformation was adopted for the statistical analysis. Note that in practice, the most common transformation understood by clinicians is the log transformation that we used for the classification method.

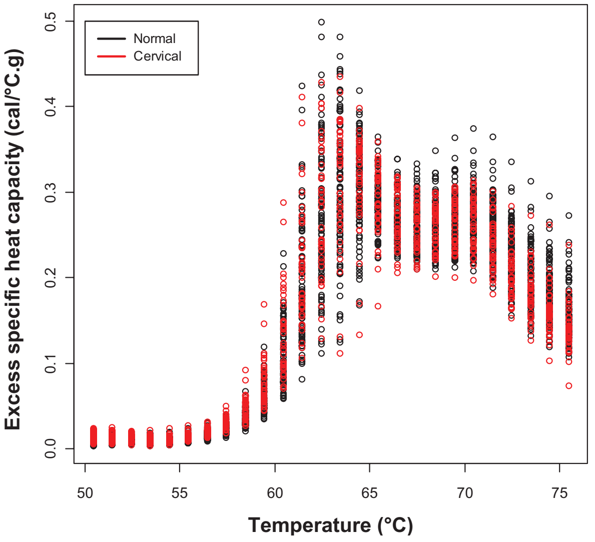

Figure 3 Scatter plot of DSC thermogram data for the Normal and Cervical groups, after data reduction.
Abbreviation: DSC, differential scanning calorimetry.

Model fitting

Using the proposed method, in Step 1, we fit the regression model (Equation 1) using all observations from 136 samples

$$H_1 = \alpha + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \beta_4 T_4 + \beta_5 G + \beta_6 G \times T_1 \\ + \beta_7 G \times T_3 + \beta_8 G \times T_4 + \varepsilon$$

where $H_1 = \log(H), T_1 = T - 62.5/25, T_i = T_1^i$ ($i = 2, 3$ and 4), and G is the group indicator, 0 for Normal and 1 for Cervical. $G \times T_2$ was excluded from the model because of its insignificancy. The resulting coefficient estimates, standard

Table 2 *P*-values from the Shapiro–Wilk normality test at each temperature point

Temperature (°C)	Transformation					
	H	H_1	H_2	H_3	H_4	H_5
50	0.000	0.113	0.097	0.097	0.000	0.000
51	0.000	0.447	0.438	0.438	0.000	0.000
52	0.000	0.562	0.568	0.568	0.000	0.000
53	0.000	0.896	0.895	0.895	0.000	0.000
54	0.000	0.381	0.345	0.345	0.000	0.000
55	0.000	0.847	0.830	0.830	0.000	0.000
56	0.003	0.636	0.631	0.631	0.003	0.003
57	0.000	0.595	0.544	0.544	0.000	0.000
58	0.000	0.089	0.039	0.039	0.000	0.000
59	0.000	0.025	0.003	0.003	0.000	0.000
60	0.000	0.011	0.000	0.000	0.000	0.000
61	0.000	0.312	0.000	0.000	0.000	0.000
62	0.090	0.136	0.000	0.000	0.116	0.209
63	0.357	0.000	0.106	0.090	0.281	0.118
64	0.000	0.000	0.009	0.010	0.000	0.000
65	0.052	0.000	0.083	0.085	0.043	0.024
66	0.306	0.590	0.248	0.247	0.322	0.369
67	0.053	0.290	0.045	0.045	0.057	0.071
68	0.061	0.538	0.036	0.036	0.069	0.094
69	0.186	0.799	0.075	0.073	0.207	0.276
70	0.247	0.677	0.080	0.079	0.273	0.350
71	0.193	0.671	0.103	0.101	0.215	0.285
72	0.133	0.879	0.231	0.228	0.150	0.208
73	0.011	0.931	0.096	0.095	0.013	0.021
74	0.000	0.479	0.011	0.011	0.000	0.001
75	0.000	0.010	0.001	0.001	0.000	0.000
Mean	0.065	0.420	0.212	0.211	0.067	0.078
No of $P > 0.05$	10	20	16	16	9	9
% of $P > 0.05$	38	77	62	62	35	35

Note: H is the heat capacity.

errors and *P*-values are presented in Table 3. Having defined the parameters of the classification model, the quantiles $q_1, \ldots, q_{26}$ were subsequently derived. We then randomly selected 57 out of the 97 Normal samples and 19 from the 39 Cervical samples and used these 76 samples to fit the regression model, as described in Step 2. Following Steps 3 and 4, the model was subsequently validated using the remaining 40 Normal and 20 Cervical samples. The results are shown in Table 4. The method achieved the correct classification of 38 samples from the Normal group and 20 from the Cervical group, corresponding to a classification rate of 96.7%, (20+38)/60. The diagnostic performance was also evaluated through the calculation of Sens, Spec, PPV, and NPV with values of Sens = 1.000, Spec = 0.950, PPV = 0.909, and NPV = 1.000 (Tables 3 and 5). (Tables 2 and 6). We observed that all 120 *P*-values ($\overline{P}(G=0)$ and $\overline{P}(G=1)$ were between 0.019 and 0.237.

Table 3 Coefficients used in the regression model

Parameter	Variable	Estimator	SE	*P*-value
α	Intercept	−1.672	0.010	<0.031
β_1	T_1	6.773	0.048	
β_2	T_2	−18.332	0.212	
β_3	T_3	−16.610	0.281	
β_4	T_4	51.221	0.901	
β_5	G	0.085	0.013	
β_6	GT_1	–0.362	0.090	
β_7	GT_3	−1.118	0.516	
β_8	GT_4	2.058	0.569	
R^2	0.957			

Abbreviation: SE, standard error.

Note: $T_1 = (T - 62.5)/25$, and $T_i = T_1^i$ where (i = 2,3,4), and G is the group indicator.

Using I-RELIEF

For comparison purpose, the I-RELIEF method was employed to analyze the same data for Normal and Cervical samples and compared with the performance of our classification method described previously. The results are presented in Tables 5 and 6. The I-RELIEF method correctly classified 34 samples from the Normal group and ten from the Cervical group, to give a classification rate of 73.3%, (10 + 34)/60. This can be compared with the classification rate of 96.7% from the application of our classification method. The diagnostic performance of I-RELIEF was characterized by values of Sens = 0.500, Spec = 0.850, PPV = 0.625, and NPV = 0.773, which were significantly lower than those of Sens = 1.000, Spec = 0.950, PPV = 0.909, and NPV = 1.000, using our classification method.

Discussion

One of the focuses of our research is to reduce the dimensionality and develop a method of classification. For each subject, we have raw data from 902 repeat measures. We have reduced data dimension by systematic grouping. For the raw data, normality was often rejected, which was not the case for the grouped data. Nevertheless, our classification method has worked really well on the grouped data. So we built the model only on the reduced data. If the classification method does not work on the reduced data in another setting, one should consider different groupings or no grouping at all.

Table 4 Results from the application of our method to the classification of plasma thermogram data for Normal and Cervical samples

Group	True		Total
	Case	Control	
Classified			
Case	20	2	22
Control	0	38	38
Total	20	40	60

Table 5 Results from the application of I-RELIEF (Sun and Li)[3] to the classification of plasma thermogram data for Normal and Cervical samples

Group	True		Total
	Case	Control	
Classified			
Case	10	6	16
Control	10	34	44
Total	20	40	60

We have described the development of a new method for the classification of high-dimensional data sets. The method employs sampling without replacement in combination with data splitting, to form many nonoverlapping training (model building) and validation (class prediction) sets. (The validation sets do overlap with each other, as do the training sets). In data splitting, a portion of the data is reserved for model fitting/building, and the remainder of the data set is used for outcome prediction (classification) and validation. The data splitting, that is, steps 2 and 3 in the section "The Proposed Method," is repeated 5000 times using random sampling without replacement. One exception to the data-splitting approach is the model-fitting step to the complete data set (Equation 1). However, this step is simply used to calculate a quantile to determine and compare the proportion of absolute value residuals P, which exceed the given constant for a defined class model, ie, G = 1 (cases) versus G = 0 (controls). Here, a unit in the test set is allocated to the class with the lower proportion P of residuals, which exceeds this constant. This proportion P is a dissimilarity distance, since a higher value of P is associated with higher absolute residual values and therefore a poorer predicted fit for a model assuming a given class. This approach contrasts with the I-RELIEF method, which uses a similarity distance, a weighted distance between the nearest hit and nearest miss. Moreover, whereas the I-RELIEF method is an interesting adaptation of a nonparametric method (nearest neighbors), the new method is a parametric method based on a linear model. In the latter case the residuals are used to develop the distance measure. The fact that the proposed method has a lower classification error than the I-RELIEF method may indicate that for the current data set, the linear model assumptions hold at least approximately. For further development of this classification model and for its prospective application, Equation 1 could be limited to the training set, since it is only used to determine a constant that can be used to compare the proportion of residuals exceeding the given constant, for each of the defined class models.

Table 6 Comparison between our classification method and I-RELIEF (Sun and Li)[3]

Statistic	Mean (95% CI)	
	Our method (%)	I-RELIEF (%)
Classification rate	96.7 (92.1, 100)	73.3 (62.1, 84.5)
Sensitivity	100.0 (99.6, 100)	50.0 (28.1, 71.9)
Specificity	95.0 (88.2, 100)	85.0 (73.9, 96.1)
PPV	90.9 (78.9, 100)	62.5 (38.8, 86.2)
NPV	100.0 (99.7, 100)	77.3 (64.9, 89.7)

Abbreviations: PPV, positive predictive values; NPV, negative predictive values.

Due to the high-dimensional data and varying shapes, it is somewhat cumbersome to build any simplistic model, such as auto-regressive of lag 1, for correlation in simulations, to study properties of the proposed and existing methods of classification. However, we plan to consider some parametric models as extensions of this research.

Conclusion

This report proposes a new classification method for the analysis of high-dimensional data sets. The method is simple but efficient and relies entirely on statistical analysis. We have demonstrated that our method performs significantly better than I-RELIEF for the application of DSC thermogram data. Future direction includes evaluation of the new classification method with additional data sets and modification of the method to include use of only training set data for the model-fitting step and a mixed-model approach to account for between-subject variation.

Acknowledgments

This work was supported by a grant from the Elsa U Pardee Foundation (to JBC) and by the James Graham Brown Foundation. SNR and JP were partly supported by Dr Donald Miller and the Wendell Cherry Chair in Clinical Trial Research. We sincerely thank reviewers and the editor for their insightful comments, which helped improve the quality of research.

Disclosure

NCG and JBC are coinventors on patent applications describing the DSC plasma thermogram technology for which Louisville Bioscience, Inc holds an exclusive license from the University of Louisville. JBC is a founder and shareholder of Louisville Bioscience, Inc; NCG is a founder, shareholder, and employee of Louisville Bioscience, Inc.

References

1. Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*. 2007;23(1):30–37.

2. Kira K, Rendell LA. A practical approach to feature selection, In Proceedings of the 9th International Workshop on Machine Learning; 1992; Aberdeen, UK. San Francisco: Morgan Kaufmann Publishers Inc; 1992: 249–256.
3. Sun Y, Li J. Iterative RELIEF for feature weighting. Proceedings of the 23rd International Conference on Machine Learning; June 25–29, 2006; Pittsburgh, USA. New York: Association for Computing Machinery (ACM); 2006. Available from: http://dl.acm.org/citation.cfm?id=1143959&dl=ACM&coll=DL&CFID=160989306&CFTOKEN=66091717. Accessed December 22, 2012.
4. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data Anal*. 2005;48:869–885.
5. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
6. Vapnik V. *Statistical Learning Theory*. Chichester: Wiley; 1998.
7. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1): 119–139.
8. Dietterich TG. Machine learning research: four current directions. *AI Magazine*. 1997;18:97–136.
9. Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. *Biostatistics*. 2009;10(1):94–105.
10. Garbett NC, Miller JJ, Jenson AB, Chaires JB. Calorimetry outside the box: a new window into the plasma proteome. *Biophys J*. 2008;94(4): 1377–1383.
11. Garbett NC, Miller JJ, Jenson AB, Chaires JB. Calorimetric analysis of the plasma proteome. *Semin Nephrol*. 2007;27(6):621–626.
12. Garbett NC, Miller JJ, Jenson AB, Miller DM, Chaires JB. Interrogation of the plasma proteome with differential scanning calorimetry. *Clin Chem*. 2007;53(11):2012–2014.
13. Garbett NC, Mekmaysy CS, Helm CW, Jenson AB, Chaires JB. Differential scanning calorimetry of blood plasma for clinical diagnosis and monitoring. *Exp Mol Pathol*. 2009;86(3):186–191.