Annals of Biomedical Engineering, Vol. 35, No. 6, June 2007 (*© 2007*) pp. 1095–1109
DOI: 10.1007/s10439-007-9317-7

Large-Scale Optimization-Based Classification Models in Medicine and Biology

EVA K. LEE [1,2,3]

[1]Center for Operations Research in Medicine and HealthCare, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA; [2]Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA; and [3]Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA 30322, USA

(Received 29 December 2006; accepted 11 April 2007; published online 15 May 2007)

Abstract—We present novel optimization-based classification models that are general purpose and suitable for developing predictive rules for large heterogeneous biological and medical data sets. Our predictive model simultaneously incorporates (1) the ability to classify any number of distinct groups; (2) the ability to incorporate heterogeneous types of attributes as input; (3) a high-dimensional data transformation that eliminates noise and errors in biological data; (4) the ability to incorporate constraints to limit the rate of misclassification, and a reserved-judgment region that provides a safeguard against over-training (which tends to lead to high misclassification rates from the resulting predictive rule); and (5) successive multi-stage classification capability to handle data points placed in the reserved-judgment region. To illustrate the power and flexibility of the classification model and solution engine, and its multi-group prediction capability, application of the predictive model to a broad class of biological and medical problems is described. Applications include: the differential diagnosis of the type of erythemato-squamous diseases; predicting presence/absence of heart disease; genomic analysis and prediction of aberrant CpG island meythlation in human cancer; discriminant analysis of motility and morphology data in human lung carcinoma; prediction of ultrasonic cell disruption for drug delivery; identification of tumor shape and volume in treatment of sarcoma; discriminant analysis of biomarkers for prediction of early atherosclerois; fingerprinting of native and angiogenic microvascular networks for early diagnosis of diabetes, aging, macular degeneracy and tumor metastasis; prediction of protein localization sites; and pattern recognition of satellite images in classification of soil types. In all these applications, the predictive model yields correct classification rates ranging from 80 to 100%. This provides motivation for pursuing its use as a medical diagnostic, monitoring and decision-making tool.

Keywords—Prediction, Predictive health, Discriminant analysis, Machine learning, Discrete support-vector machine, Multi-category classification models, Integer programming, Optimization, Reserved judgment, Multi-stage classification, Medical diagnosis.

INTRODUCTION

A fundamental problem in discriminant analysis, or supervised learning, concerns the classification of an entity into one of $G(G \geq 2)$*a priori*, mutually exclusive groups based upon k specific measurable features of the entity. Typically, a discriminant rule is formed from data collected on a sample of entities for which the group classifications are known. Then new entities, whose classifications are unknown, can be classified based on this rule. Such an approach has been applied in a variety of domains, and a large body of literature on both the theory and applications of discriminant-analysis exists (e.g., see the bibliography in McLachlan[67]).

In experimental biological and medical research, very often, experiments are performed and measurements are recorded under different conditions and/or on different cells/molecules. A critical analysis involves the discrimination of different features under different conditions that will reveal potential predictors for biological and medical phenomena. Hence, classification techniques play an extremely important role in biological analysis, as they facilitate systematic correlation and classification of different biological and medical phenomena. A resulting predictive rule can assist, for example, in early disease prediction and diagnosis, identification of new target sites (genomic, cellular, molecular) for treatment and drug delivery, disease prevention and early intervention, and optimal treatment design.

Address correspondence to Eva K. Lee, Center for Operations Research in Medicine and HealthCare, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA. Electronic mail: eva.lee@isye.gatech.edu

There are five fundamental steps in discriminant analysis. (a) Determine the data for input and the predictive output classes. (b) Gather a training set of data (including output class) from human experts or from laboratory experiments. Each element in the training set is an entity with corresponding known output class. (c) Determine the input attributes to represent each entity. (d) Identify discriminatory attributes and develop the predictive rule(s); (e) Validate the performance of the predictive rule(s).

In our Center for Operations Research in Medicine and HealthCare, we have developed a general-purpose discriminant analysis modeling framework and computational engine for various biological and biomedical informatics analyses. Our model, the first discrete support-vector machine, offers distinct features (e.g., the ability to classify any number of groups, management of the *curse of dimensionality* in data attributes, and a reserved-judgment region to facilitate multi-stage classification analysis) that are not simultaneously available in existing classification software.[32,33,49,50,56] Studies involving tumor volume identification, ultrasonic cell disruption in drug delivery, lung tumor cell motility analysis, CpG island aberrant methylation in human cancer, predicting early atherosclerosis using biomarkers, and fingerprinting native and angiogenic microvascular networks using functional perfusion data indicate that our zapproach is adaptable and can produce effective and reliable predictive rules for various biomedical and bio-behavior phenomena.[13,25,26,51,53,55,57],

Section "Background" briefly describes the background of discriminant analysis. Section "Discrete Support-Vector Machine Predictive Models" describes the optimization-based multi-stage discriminant analysis predictive models for classification. The use of the predictive models on various biological and medical problems are presented in Section "Classification Results on Real-World Applications." This is followed by a brief summary in Section "Summary and Conclusion."

BACKGROUND

The main objective in discriminant analysis is to derive rules that can be used to classify entities into groups. Discriminant rules are typically expressed in terms of variables representing a set of measurable attributes of the entities in question. Data on a sample of entities for which the group classifications are known (perhaps determined by extraordinary means) are collected and used to derive rules that can be used to classify new yet-to-be-classified entities. Often there is a trade-off between the discriminating ability of the selected attributes and the expense of obtaining measurements on these attributes. Indeed, the measurement of a relatively definitive discriminating feature may be prohibitively expensive to obtain on a routine basis, or perhaps impossible to obtain at the time that classification is needed.

Thus, a discriminant rule based on a selected set of feature attributes will typically be an imperfect discriminator, sometimes misclassifying entities. Depending on the application, the consequences of misclassifying an entity may be substantial. In such a case, it may be desirable to form a discrimination rule that allows less specific classification decisions, or even non-classification of some entities to reduce the probability of misclassification.

To address this concern, a number of researchers have suggested methods for deriving *partial discrimination rules*.[12,37,41,72,75] A partial discrimination rule allows an entity to be classified into some subset of the groups (i.e., rule out membership in the remaining groups), or be placed in a "reserved-judgment" category. An entity is considered misclassified only when it is assigned to a non-empty subset of groups not containing the true group of the entity. Typically, methods for deriving partial discrimination rules attempt to constrain the misclassification probabilities (e.g., by enforcing an upper bound on the proportion of misclassified training sample entities). For this reason, the resulting rules are also sometimes called *constrained discrimination rules.*

Partial (or constrained) discrimination rules are intuitively appealing. A partial discrimination rule based on relatively inexpensive measurements can be tried first. If the rule classifies the entity satisfactorily according to the needs of the application, then nothing further needs to be done. Otherwise, additional measurements—albeit more expensive—can be taken on other, more definitive, discriminating attributes of the entity.

One disadvantage of partial discrimination methods is that there is no obvious definition of optimality among any set of rules satisfying the constraints on the misclassification probabilities. For example, since some correct classifications are certainly more valuable than others (e.g., classification into a small subset containing the true group vs. a large subset), it does not make sense simply to maximize the probability of correct classification. In fact, to maximize the probability of correct classification, one would simply classify every entity into the subset consisting of all the groups—clearly, not an acceptable rule.

A simplified model, whereby one incorporates only the reserved-judgment region (i.e., an entity is either classified as belonging to exactly one of the given *a priori* groups, or it is placed in the reserved-judgment category), is amenable to reasonable notions of optimality. For example, in this case, maximizing the probability of correct classification is meaningful. For the two-group case, the simplified model and the more general model are equivalent. Research on the two-group case is summarized in McLachlan.[67] For three or more groups, the two models are not equivalent, and most work has been directed toward the development of heuristic methods for the more general model.[12,37,72,75]

Assuming that the group density functions and prior probabilities are known, Anderson[1] showed that an optimal rule for the problem of maximizing the probability of correct classification subject to constraints on the misclassification probabilities must be of a specific form when discriminating among multiple groups with a simplified model. The formulae in Anderson's result depend on a set of parameters satisfying a complex relationship between the density functions, the prior probabilities, and the bounds on the misclassification probabilities. Establishing a viable mathematical model to describe Anderson's result, and finding values for these parameters that yield an optimal rule are challenging tasks. Gallagher *et al.*[33] presented the first computational model for Anderson's results.

A variety of mathematical-programming models have been proposed for the discriminant-analysis problem.[3,5,6,16,28,29,36,38,39,40,44,61,63,64,74,80,82] None of these studies deal formally with measuring the performance of discriminant rules specifically designed to allow allocation to a reserved-judgment region. There is also no mechanism employed to constrain the level of misclassifications for each group.

Many different techniques and methodologies have contributed to advances in classification, including artificial neural networks, decision trees, kernel-based learning, machine learning, mathematical programming, statistical analysis, and support-vector machines.[7,10,21,23,62,68,84] There are some review papers for classification problems with mathematical-programming techniques. Stam[79] summarizes basic concepts and ideas and discusses potential research directions on classification methods that optimize a function of the L_p-norm distances. The paper focuses on continuous models and includes normalization schemes, computational aspects, weighted formulations, secondary criteria, and extensions from two-group to multi-group classifications. Zopounidis and Doumpos[89] review the research conducted on the framework of the multi-criteria decision aiding, covering different classification models. Mangasarian[65] and Bradley *et al.*[9] give an overview of using mathematical-programming approaches to solve data mining problems. Most recently, Lee and Wu[60] provide a comprehensive overview of continuous and discrete mathematical-programming models for classification problems.

DISCRETE SUPPORT-VECTOR MACHINE PREDICTIVE MODELS

In our computational center, since 1997, we have been developing a general-purpose discriminant-analysis modeling framework and computational engine that is applicable to a wide variety of applications, including biological, biomedical and logistics problems. Utilizing the technology of large-scale discrete optimization and support-vector machines, we have developed novel predictive models that simultaneously include the following features: (1) the ability to classify any number of distinct groups; (2) the ability to incorporate heterogeneous types of attributes as input; (3) a high-dimensional data transformation that eliminates noise and errors in biological data; (4) constraints to limit the rate of misclassification, and a reserved-judgment region that provides a safeguard against over-training (which tends to lead to high misclassification rates from the resulting predictive rule); and (5) successive multi-stage classification capability to handle data points placed in the reserved-judgment region. Based on the description in Gallagher *et al.*,[32,33] Lee *et al.*,[56] and Lee,[49,50] we summarize below some of the classification models we have developed.

Modeling of Reserved-Judgment Region for General Groups

When the population densities and prior probabilities are known, the constrained rules with a reject option (reserved judgment), based on Anderson's results, calls for finding a partition $\{R_0, \ldots, R_G\}$ of $\Re^k$ that maximizes the probability of correct allocation subject to constraints on the misclassification probabilities; i.e.

$$\text{Maximize} \sum_{g=1}^{G} \pi_g \int_{R_g} f_g(w)dw \tag{1}$$

$$\text{Subject to} \int_{R_g} f_h(w)dw \leq \alpha_{hg}, \quad h, g = 1, \ldots, G, \ h \neq g, \tag{2}$$

where $f_h, h = 1, \ldots, G$, are the group conditional density functions, π_g denotes the prior probability that a randomly selected entity is from group $g, g = 1, \ldots, G$, and $\alpha_{hg}, h \neq g$, are constants between zero and one.

Under quite general assumptions, it was shown that there exist unique (up to a set of measure zero) non-negative constants $\lambda_{ih}, i, h \in \{1, \ldots, G\}, i \neq h$, such that the optimal rule is given by

$$R_g = \{x \in \Re^k : L_g(x) = \max_{h \in \{0,1,\ldots G\}} L_h(x)\}, \quad g = 0, \ldots, G \tag{3}$$

where

$$L_0(x) = 0 \tag{4}$$

$$L_h(x) = \pi_h f_h(x) - \sum_{\substack{i=1 \\ i \neq h}}^{G} \lambda_{ih} f_i(x), \quad h = 1, \ldots, G \tag{5}$$

For $G = 2$ the optimal solution can be modeled rather straightforward. However, finding optimal λ_{ih}s for the general case, $G \geq 3$, is a difficult problem, with the difficulty increasing as G increases. Our model offers an avenue for modeling and finding the optimal solution in the general case. It is the first such model to be computationally viable.[32,33]

Before proceeding, we note that R_g can be written as $R_g = \{x \in \Re^k : L_g(x) \geq L_h(x) \text{ for all } h = 0, \ldots, G\}$. So, since $L_g(x) \geq L_h(x)$ if, and only if, $(1/\sum_{t=1}^{G} f_t(x)) L_g(x) \geq (1/\sum_{t=1}^{G} f_t(x)) L_h(x)$, the functions $L_h, h = 1, \ldots, G$, can be redefined as

$$L_h(x) = \pi_h p_h(x) - \sum_{\substack{i=1 \\ i \neq h}}^{G} \lambda_{ih} p_i(x) \quad h = 1, \ldots, G \tag{6}$$

where $p_i(x) = f_i(x)/\sum_{t=1}^{G} f_t(x)$. We assume that L_h is defined as in Eq. (6) in our model.

Mixed Integer Programming (MIP) Formulations

Assume that we are given a training sample of N entities whose group classifications are known; say n_g entities are in group g, where $\sum_{g=1}^{G} n_g = N$. Let the k dimensional vectors $x^{gj}, g = 1, \ldots, G, j = 1, \ldots, n_g$, contain the measurements on k available characteristics of the entities. Our procedure for deriving a discriminant rule proceeds in two stages. The first stage is to use the training sample to compute estimates, $\hat{f}_h$, either parametrically or non-parametrically, of the density functions f_h[67] and estimates, $\hat{\pi}_h$, of the prior probabilities $\pi_h, h = 1, \ldots, G$. The second stage is to determine the optimal λ_{ih}s given these estimates. This stage requires being able to estimate the probabilities of correct classification and misclassification for any candidate set of λ_{ih}s. One could, in theory, substitute the estimated densities and prior probabilities into equations (5), and directly use the resulting regions R_g in the integral expressions given in (1) and (2). This would involve, even in simple cases such as normally distributed groups, the numerical evaluation of k-dimensional integrals at each step of a search for the optimal λ_{ih}s. Therefore, we have designed an alternative approach. After substituting the $\hat{f}_h$'s and $\hat{\pi}_h$'s into Eq. (5), we simply calculate the proportion of training sample points which fall in each of the regions $R_1, \ldots, R_G$. The MIP models discussed below attempt to maximize the proportion of training sample points correctly classified while satisfying constraints on the proportions of training sample points misclassified. This approach has two advantages. First, it avoids having to evaluate the potentially difficult integrals in Eqs. (1) and (2). Second, it is non-parametric in controlling the training sample misclassification probabilities. That is, even if the densities are poorly estimated (by assuming, for example, normal densities for non-normal data), the constraints are still satisfied for the training sample. Better estimates of the densities may allow a higher correct classification rate to be achieved, but the constraints will be satisfied even if poor estimates are used. Unlike most support-vector machine models that minimize the sum of errors, our objective is driven by the number of correct classifications, and will not be biased by the distance of the entities from the supporting hyperplane.

A word of caution is in order. In traditional unconstrained discriminant analysis, the true probability of correct classification of a given discriminant rule tends to be smaller than the rate of correct classification for the training sample from which it was derived. One would expect to observe such an effect for the method described herein as well. In addition, one would expect to observe an analogous effect with regard to constraints on misclassification probabilities—the true probabilities are likely to be greater than any limits imposed on the proportions of training sample misclassifications. Hence, the α_{hg} parameters should be carefully chosen for the application in hand.

Our first model is a non-linear 0/1 MIP model with the non-linearity appearing in the constraints. Model 1 maximizes the number of correct classifications of the given N training entities. Similarly, the constraints on the misclassification probabilities are modeled by ensuring that the number of group g training entities in region R_h is less than or equal to a pre-specified percentage, $\alpha_{hg} (0 < \alpha_{hg} < 1)$, of the total number, n_g, of group g entities, $h, g \in \{1, \ldots, G\}, \text{h} \neq \text{g}$.

For notational convenience, let $\mathbf{G} = \{1, \ldots, G\}$ and $\mathbf{N}_\text{g} = \{1, \ldots, n_g\}$, for $g \in \mathbf{G}$. Also, analogous to the definition of p_i, define $\hat{p}_i$ by $\hat{p}_i(x) = \hat{f}_i(x)/\sum_{t=1}^{G} \hat{f}_t(x)$.

In our model, we use binary indicator variables to denote the group classification of entities. Mathematically, let u_{hgj} be a binary variable indicating whether or not x^{gj} lies in region R_h; i.e., whether or not the jth entity from group g is allocated to group h. Then Model 1 can be written as follows:

$$\text{Maximize} \sum_{g \in G} \sum_{j \in N_g} u_{ggj}$$

Subject to

$$L_{hgj} = \hat{\pi}_h \hat{p}_h(x^{gj}) - \sum_{i \in G \backslash h} \lambda_{ih} \hat{p}_i(x^{gj}) \quad h, g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{7}$$

$$y_{gj} = \max\{0, L_{hgj} : h = 1, \ldots, G\} \quad g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{8}$$

$$y_{gj} - L_{ggj} \leq M(1 - u_{ggj}) \quad g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{9}$$

$$y_{gj} - L_{hgj} \geq \varepsilon(1 - u_{hgj}) \quad h, g \in \mathbf{G},\ j \in \mathbf{N_g}, h \neq g \tag{10}$$

$$\sum_{j \in N_g} u_{hgj} \leq \lfloor \alpha_{hg} n_g \rfloor \quad h, g \in \mathbf{G},\ h \neq g \tag{11}$$

$$-\infty < L_{hgj} < \infty, \quad y_{gj} \geq 0, \quad \lambda_{ih} \geq 0, \quad u_{hgj} \in \{0, 1\}$$

Constraint (7) defines the variable L_{hgj} as the value of the function L_h evaluated at x^{gj}. Therefore, the continuous variable y_{gj}, defined in constraint (8), represents $\max\{L_h(x^{gj}) : h = 0, \ldots, G\}$; and consequently, x^{gj} lies in region R_h if, and only if, $y_{gj} = L_{hgj}$. The binary variable u_{hgj} is used to indicate whether or not x^{gj} lies in region R_h; i.e., whether or not the jth entity from group g is allocated to group h. In particular, constraint (9), together with the objective, force u_{ggj} to be 1 if, and only if, the jth entity from group g is correctly allocated to group g; and constraints (10) and (11) ensure that at most $\lfloor \alpha_{hg} n_g \rfloor$ (i.e., the greatest integer less than or equal to $\alpha_{hg} n_g$) group g entities are allocated to group $h, h \neq g$. One caveat regarding the indicator variables u_{hgj} is that although the condition $u_{hgj} = 0, h \neq g$, implies (by constraint (10)) that $x^{gj} \notin R_h$, the converse need not hold. As a consequence, the number of misclassifications may be overcounted. However, in our preliminary numerical study we found that the actual amount of overcounting is minimal. One could force the converse (thus, $u_{hgj} = 1$ if and only if $x^{gj} \in R_h$) by adding constraints $y_{gj} - L_{hgj} \leq M(1 - u_{hgj})$, for example. Finally, we note that the parameters M and ε are extraneous to the discriminant-analysis problem itself, but are needed in the model to control the indicator variables u_{hgj}. The intention is for M and ε to be, respectively, large and small positive constants.

Model Variations

We explore different variations in the model to grasp the quality of the solution and the associated computational effort.

A first variation involves transforming Model 1 to an equivalent linear mixed integer model. In particular, Model 2 replaces the N constraints defined in (8) with the following system of $3GN + 2N$ constraints:

$$y_{gj} \geq L_{hgj} \quad h, g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{12}$$

$$\tilde{y}_{hgj} - L_{hgj} \leq M(1 - v_{hgj}) \quad h, g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{13}$$

$$\tilde{y}_{hgj} \leq \hat{\pi}_h \hat{p}_h(x^{gj}) v_{hgj} \quad h, g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{14}$$

$$\sum_{h \in \mathcal{G}} v_{hgj} \leq 1 \quad g \in ,\ j \in \mathbf{N_g} \tag{15}$$

$$\sum_{h \in \mathcal{G}} \tilde{y}_{hgj} = y_{gj} \quad g \in \mathbf{G},\ j \in \mathbf{N_g} \tag{16}$$

where $\tilde{y}_{hgj} \geq 0$ and $v_{hgj} \in \{0, 1\}, h, g \in \mathbf{G}, j \in \mathbf{N_g}$. These constraints, together with the non-negativity of y_{gj} force $y_{gj} = \max\{0, L_{hgj} : h = 1, \ldots, G\}$.

The second variation involves transforming Model 1 to a heuristic linear MIP model. This is done by replacing the non-linear constraint (8) with $y_{gj} \geq L_{hgj}, h, g \in \mathbf{G}, j \in \mathbf{N_g}$, and including penalty terms in the objective function. In particular, Model 3 has the objective

$$\text{Maximize} \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} \beta u_{ggj} - \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} \gamma y_{gj},$$

where β and γ are positive constants. This model is heuristic in that there is nothing to force $y_{gj} = \max\{0, L_{hgj} : h = 1, \ldots, G\}$. However, since in addition to trying to force as many u_{ggj}'s to one as possible, the objective in Model 3 also tries to make the y_{gj}'s as small as possible, and the optimizer tends to drive y_{gj} toward $\max\{0, L_{hgj} : h = 1, \ldots, G\}$.We remark that β and γ could be stratified by group (i.e., introduce possibly distinct $\beta_g, \gamma_g, g \in \mathbf{G}$) to model the relative importance of certain groups to be correctly classified.

A reasonable modification to Models 1, 2, and 3 involves relaxing the constraints specified by (11). Rather than placing restrictions on the number of type g training entities classified into group h, for all $h, g \in \mathbf{G}, h \neq g$, one could simply place an upper bound on the *total* number of misclassified training entities. In this case, the $G(G-1)$ constraints specified by (11) would be replaced by the single constraint

$$\sum_{g \in G} \sum_{h \in G \backslash \{g\}} \sum_{j \in N_g} u_{hgj} \leq \lfloor \alpha N \rfloor \tag{17}$$

where α is a constant between 0 and 1. We will refer to Models 1, 2, and 3, modified in this way, as Models 1T, 2T, and 3T, respectively. Of course, other modifications are also possible. For instance, one could place restrictions on the total number of type g points misclassified for each $g \in \mathbf{G}$.Thus, in place of the constraints specified in (17), one would include the constraints $\sum_{h \in G\backslash\{g\}} \sum_{j \in N_g} u_{hgj} \leq \lfloor \alpha_g N \rfloor, g \in \mathbf{G}$, where $0 < \alpha_g < 1$.

We also explore a heuristic linear model of Model 1. In particular, consider the linear program (DALP):

$$\text{Minimize} \sum_{g \in G} \sum_{j \in N_g} \left(c_1 w_{gj} + c_2 y_{gj}\right) \qquad (18)$$

Subject to

$$L_{hgj} = \pi_h \hat{p}_h(x^{gj}) - \sum_{i \in G\backslash\{h\}} \lambda_{ih} \hat{p}_i(x^{gj}) \quad h, g \in \mathbf{G},\ j \in \mathbf{N}_g \qquad (19)$$

$$L_{ggj} - L_{hgj} + w_{gj} \geq 0 \quad h, g \in \mathbf{G},\ h \neq g,\ j \in \mathbf{N}_g \quad (20)$$

$$L_{ggj} + w_{gj} \geq 0 \quad g \in \mathbf{G},\ j \in \mathbf{N}_g \qquad (21)$$

$$-L_{hgj} + y_{gj} \geq 0 \quad h, g \in \mathbf{G},\ j \in \mathbf{N}_g \qquad (22)$$

$$-\infty < L_{hgj} < \infty,\ w_{gj}, y_{gj},\ \ \lambda_{ih} \geq 0$$

Constraint (19) defines the variable L_{hgj} as the value of the function L_h evaluated at x^{gj}. As the optimization solver searches through the set of feasible solutions, the λ_{ih} variables will vary, causing the L_{hgj} variables to assume different values. Constraints (20), (21), and (22) link the objective-function variables with the L_{hgj} variables in such a way that correct classification of training entities, and allocation of training entities into the reserved-judgment region, are captured by the objective-function variables. In particular, if the optimization solver drives w_{gj} to zero for some g, j pair, then constraints (20) and (21) imply that $L_{ggj} = \max\{0, L_{hgj} : h \in \mathbf{G}\}$. Hence, the jth entity from group g is correctly classified. If, on the other hand, the optimal solution yields $y_{gj} = 0$ for some g, j pair, then constraint (22) implies that $\max\{0, L_{hgj} : h \in \mathbf{G}\} = 0$. Thus, the jth entity from group g is placed in the reserved-judgment region. (Of course, it is possible for both w_{gj} and y_{gj} to be zero. One should decide prior to solving the linear program how to interpret the classification in such cases.) If both w_{gj} and y_{gj} are positive, the j th entity from group g is misclassified.

The optimal solution yields a set of λ_{ih}s that best allocates the training entities (i.e., "best" in terms of minimizing the penalty objective function). The optimal λ_{ih}s can then be used to define the functions $L_h, h \in \mathbf{G}$,which in turn can be used to classify a new entity with feature vector $x \in \Re^k$ by simply computing the index at which $\max\{L_h(x) : h \in \{0, 1, \ldots, G\}\}$ is achieved.

Note that Model DALP places no *a priori* bound on the number of misclassified training entities. However, since the objective is to minimize a weighted combination of the variables w_{gj} and y_{gj}, the optimizer will attempt to drive these variables to zero. Thus, the optimizer is, in essence, attempting either to correctly classify training entities ($w_{gj} = 0$), or to place them in the reserved-judgment region ($y_{gj} = 0$). By varying the weights c_1 and c_2, one has a means of controlling the optimizer's emphasis for correctly classifying training entities vs. placing them in the reserved-judgment region. If $c_2/c_1 < 1$, the optimizer will tend to place a greater emphasis on driving the w_{gj} variables to zero than driving the y_{gj} variables to zero (conversely, if $c_2/c_1 > 1$). Hence, when $c_2/c_1 < 1$, one should expect to get relatively more entities correctly classified, fewer placed in the reserved-judgment region, and more misclassified, than when $c_2/c_1 > 1$. An extreme case is when $c_2 = 0$. In this case, there is no emphasis on driving y_{gj} to zero (the reserved-judgment region is thus ignored), and the full emphasis of the optimizer is to drive w_{gj} to zero.

Table 1 summarizes the number of constraints, the total number of variables, and the number of 0/1 variables in each of the discrete support-vector machine models, and in the heuristic LP model

TABLE 1. Model size.

Model	Type	Constraints	Total variables	0/1 Variables
1	Non-linear MIP	$2GN+N+G(G-1)$	$2GN+N+G(G-1)$	GN
2	Linear MIP	$5GN+2N+G(G-1)$	$4GN+N+G(G-1)$	$2GN$
3	Linear MIP	$3GN+G(G-1)$	$2GN+N+G(G-1)$	GN
1T	Non-linear MIP	$2GN+N+1$	$2GN+N+G(G-1)$	GN
2T	Linear MIP	$5GN+2N+1$	$4GN+N+G(G-1)$	$2GN$
3T	Linear MIP	$3GN+1$	$2GN+N+G(G-1)$	GN
DALP	Linear program	$3GN$	$NG+N+G(G-1)$	0

(DALP). Clearly, even for moderately sized discriminant-analysis problems, the MIP instances are relatively large. Also, note that Model 2 is larger than Model 3, both in terms of the number of constraints and the number of variables. However, it is important to keep in mind that the difficulty of solving an MIP problem cannot, in general, be predicted solely by its size; problem structure has a direct and substantial bearing on the effort required to find optimal solutions. The LP relaxation of these MIP models pose computational challenges as commercial LP solvers return (optimal) LP solutions that are infeasible, due to the equality constraints, and the use of big M and small ε in the formulation.

It is interesting to note that the set of feasible solutions for Model 2 is "tighter" than that for Model 3. In particular, if F_i denotes the set of feasible solutions of Model i, then

$$\begin{aligned} F_1 = \{ & (L, \lambda, u, y) : \text{ there exists } \tilde{y}, v \\ & \text{such that } (L, \lambda, u, y, \tilde{y}, v) \in F_2 \} \} \subsetneq F_3. \end{aligned} \tag{23}$$

The novelties of the classification models developed herein include: (1) they are suitable for discriminant analysis given any number of groups, (2) they accept heterogeneous types of attributes as input, (3) they use a parametric approach to reduce high-dimensional attribute spaces, and (4) they allow constraints on the number of misclassifications, and utilize a reserved judgment to facilitate the reduction of misclassifications. The latter point opens the possibility of performing multi-stage analysis.

Clearly, the advantage of an LP model over an MIP model is that the associated problem instances are computationally much easier to solve. However, the most important criterion in judging a method for obtaining discriminant rules is how the rules perform in correctly classifying new unseen entities. Once the rule is developed, applying it to a new entity to determine its group is trivial. Extensive computational experiments have been performed to gauge the qualities of solutions of different models.[14,33,49,50,56]

Computational Strategies

The MIP models described herein offer a computational avenue for numerically estimating optimal values for the λ_{ih} parameters in Anderson's formulae. However, it should be emphasized that MIP problems are themselves difficult to solve. (Anderson[1] himself noted the extreme difficulty of finding an optimal set of λ_{ih}s.) Indeed, MIP is an NP-hard problem.[35] Nevertheless, due to the fact that integer variables—and in particular, 0/1 variables—are a powerful modeling tool, a wide variety of real-world problems have been modeled as mixed integer programs. Consequently, much effort has been invested in developing computational strategies for solving MIP problem instances.

The numerical work reported in Section "Classification Results on Real-World Applications" is based on an MIP solver which is built on top of a general-purpose mixed integer research code, MIPSOL.[45] (A competitive commercial solver (CPLEX) was not effective in solving the problem instances considered.) The general-purpose code integrates state-of-the-art MIP computational devices such as problem preprocessing, primal heuristics, global, and local reduced-cost fixing, and cutting planes into a branch-and-bound framework. The code has been shown to be effective in solving a wide variety of large-scale real-world instances.[8] For our MIP instances, special techniques such as variable aggregation, a heuristic branching scheme, and hypergraphic cut generations are employed.[14,24,33]

CLASSIFICATION RESULTS ON REAL-WORLD APPLICATIONS

The main objective in discriminant analysis is to derive rules that can be used to classify entities into groups. Computationally, the challenge lies in the effort expended to develop such a rule. Feasible solutions obtained from our classification models correspond to predictive rules. Empirical results[33,56] indicate that the resulting classification model instances are computationally very challenging, and even intractable by competitive commercial MIP solvers. However, the resulting predictive rules prove to be very promising, offering correct classification rates on new unknown data ranging from 80 to 100% on various types of biological/medical problems. Our results indicate that the general-purpose classification framework that we have designed has the potential to be a very powerful predictive method for clinical setting.

The choice of MIP as the underlying modeling and optimization technology for our support-vector machine classification model is guided by the desire to simultaneously incorporate a variety of important and desirable properties of predictive models within a general framework. MIP itself allows for incorporation of continuous and discrete variables, and linear and non-linear constraints, providing a flexible and powerful modeling environment.

Validation of Model and Computational Effort

We performed 10-fold cross-validation, and designed simulation and comparison studies on our preliminary models. The results, reported in Gallagher *et al.*,[33] and Lee *et al.*,[56] show the methods are promising, based on applications to both simulated data and real-application datasets from the machine learning database repository.[69] Furthermore, our methods compare well to existing methods, often producing better results when compared to other approaches such as artificial neural networks, quadratic discriminant analysis, tree classification, and other support-vector machines.

Applications to Biological and Medical Problems

Our mathematical modeling and computational algorithmic design shows great promise. The resulting predictive rules are able to produce higher rates of correct classification on new biological data (with unknown group status) compared to existing classification methods. This is partly due to the transformation of raw data via the set of constraints in (7). While most support-vector machines (Lee and Wu 2006[60]) directly determine the hyperplanes of separation using raw data, our approach transforms the raw data via a probabilistic model, before the determination of the supporting hyperplanes. Further, the separation is driven by maximizing the sum of binary variables (representing correct classification or not of entities), instead of minimizing a sum of errors (representing distances of entities from hyperplanes), as in other support-vector machines. The combination of these two strategies offers better classification capability. Noise in the transformed data is not as profound as in raw data. And the magnitudes of the errors do not skew the determination of the separating hyperplanes, as all entities have "equal" importance when correct classification is being counted. To highlight the broad applicability of our approach, in this paper, we briefly summarize the application of our predictive models and solution algorithms to nine different biological and medical problems. Most of the projects were carried out in close partnership with experimental biologists and/or clinicians. Applications to finance and other industry applications are described elsewhere.[14,15,33,56]

Determining the type of erythemato-squamous disease.[69] The differential diagnosis of erythemato-squamous diseases is an important problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The six groups are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages.

The six groups consist of 366 subjects (112, 61, 72, 49, 52, 20, respectively) with 34 clinical attributes. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope. The 34 attributes include (1) clinical attributes: erythema, scaling, definite borders, itching, koebner phenomenon, polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, family history, age; and (2) histopathological attributes: melanin incontinence, eosinophils in the infiltrate, PNL infiltrate, fibrosis of the papillary dermis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing of the rete ridges, elongation of the rete ridges, thinning of the suprapapillary epidermis, spongiform pustule, munro microabcess, focal hypergranulosis, disappearance of the granular layer, vacuolization and damage of basal layer, spongiosis, saw-tooth appearance of retes, follicular horn plug, perifollicular parakeratosis, inflammatory monoluclear infiltrate, band-like infiltrate.

Our multi-group classification model selected 27 discriminatory attributes, and successfully classified the patients into six groups, each with an unbiased correct classification of greater than 93% (with 100% correct rate for groups 1, 3, 5, 6) with an average overall accuracy of 98%. Using 250 subjects to develop the rule, and testing the remaining 116 patients, we obtain a prediction accuracy of 91%.

Predicting presence/absence of heart disease. The four databases concerning heart disease diagnosis were collected by Dr. Janosi of Hungarian Institute of Cardiology, Budapest; Dr. Steinbrunn of University Hospital, Zurich; Dr. Pfisterer of University Hospital, Basel, Switzerland; and Dr. Detrano of V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Each database contains the same 76 attributes.[69] The "goal" field refers to the presence of heart disease in the patient. The classification attempts to distinguish *presence* (values 1,2,3,4, involving a total of 509 subjects) from *absence*(value 0, involving 411 subjects). The attributes include demographics, physio-cardiovascular conditions, traditional risk factors, family history, personal lifestyle, and cardiovascular exercise measurements. This data set has posed some challenges to past analysis via various classification approaches, resulting in less than 80% correct classification. Applying our classification model without reserved judgment, we obtain 79% and 85% correct classification for each group respectively. To gauze the usefulness of multi-stage analysis, we apply two-stage classification. In the first stage, 14 attributes were selected as discriminatory. 135 Group *absence* subjects were placed into the reserved-judgment region, with 85% of the remaining classified as Group *absence* correctly; while 286 Group *presence* subjects were placed into the reserved-judgment region, and 91% of the remaining classified correctly into the Group *presence*. In the second stage, 11 attributes were selected with 100 and 229 classified into Group *absence* and *presence* respectively. Combining the two stages, we obtained a correct classification of 82% and 85% respectively for diagnosis of absence or presence of heart disease. Figure 1 illustrates the two-stage classification.

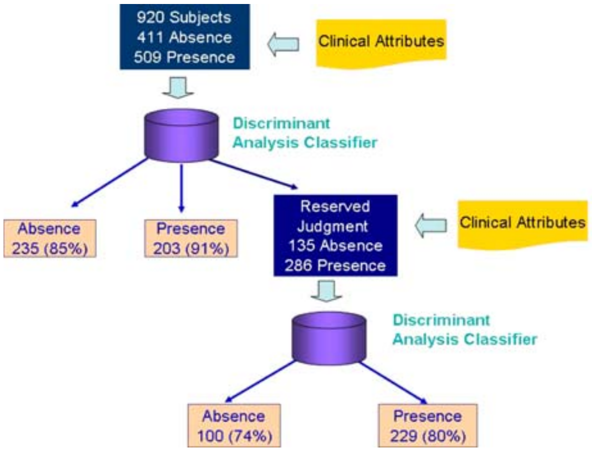

FIGURE 1. A tree diagram for two-stage classification and prediction of heart disease.

Predicting aberrant CpG Island methylation in human cancer.[25,26] Epigenetic silencing associated with aberrant methylation of promoter region CpG islands is one mechanism leading to loss of tumor suppressor function in human cancer. Profiling of CpG island methylation indicates that some genes are more frequently methylated than others, and that each tumor type is associated with a unique set of methylated genes. However, little is known about why certain genes succumb to this aberrant event. To address this question, we used Restriction Landmark Genome Scanning (RLGS) to analyze the susceptibility of 1749 unselected CpG islands to *de novo* methylation driven by overexpression of DNMT1. We found that, whereas the overall incidence of CpG island methylation was increased in cells overexpressing DNMT1, not all loci were equally affected. The majority of CpG islands (69.9%) were resistant to *de novo* methylation, regardless of DNMT1 overexpression. In contrast, we identified a subset of methylation-prone CpG islands (3.8%) that were consistently hypermethylated in multiple DNMT1 overexpressing clones. Methylation-prone and methylation-resistant CpG islands were not significantly different with respect to size, C + G content, CpG frequency, chromosomal location, or gene- or promoter-association. To discriminate methylation-prone from methylation-resistant CpG islands, we developed a novel DNA pattern recognition model and algorithm,[52] and coupled our predictive model described herein with the patterns found. We were able to derive a classification function based on the frequency of seven novel sequence patterns that was capable of discriminating methylation-prone from methylation-resistant CpG islands with 90% correctness upon cross-validation, and 85% accuracy when tested against blind CpG islands unknown to us on the methylation status. The data indicate that CpG islands differ in their intrinsic susceptibility to *de novo* methylation, and suggest that the propensity for a CpG island to become aberrantly methylated can be predicted based on its sequence context.

The significance of this research is 2-fold. First, the identification of sequence patterns/attributes that distinguish methylation-prone CpG islands will lead to a better understanding of the basic mechanisms underlying aberrant CpG island methylation. Because genes that are silenced by methylation are otherwise structurally sound, the potential for reactivating these genes by blocking or reversing the methylation process represents an exciting new molecular target for chemotherapeutic intervention. A better understanding of the factors that contribute to aberrant methylation, including the identification of sequence elements that may act to target aberrant methylation, will be an important step in achieving this long-term goal. Second, the classification of the more than 29,000 known (but as yet unclassified) CpG islands in human chromosomes will provide an important resource for the identification of novel gene targets for further study as potential molecular markers that could impact on both cancer prevention and treatment. Extensive RLGS fingerprint information (and thus potential training sets of methylated CpG islands) already exists for a number of human tumor types, including breast, brain, lung, leukemias, hepatocellular carcinomas, and PNET.[19,20,31,77] Thus, the methods and tools developed are directly applicable to CpG island methylation data derived from human tumors. Moreover, new microarray-based techniques capable of ‘profiling’ more than 7000 CpG islands have been developed and applied to human breast cancers.[11,86,87] We are uniquely poised to take advantage of the tumor CpG island methylation profile information that will likely be generated using these techniques over the next several years. Thus, our general-predictive modeling framework has the potential to lead to improved diagnosis and prognosis and treatment planning for cancer patients.

Discriminant analysis of cell motility and morphology data in human lung carcinoma.[13] This study focuses on the differential effects of extracellular matrix proteins on the motility and morphology of human lung epidermoid carcinoma cells. The behavior of carcinoma cells is contrasted with that of normal L-132 cells, resulting in a method for the prediction of metastatic potential. Data collected from time-lapsed videomicroscopy were used to simultaneously produce quantitative measures of motility and morphology. The data were subsequently analyzed using our discriminant-analysis model and algorithm to discover relationships between motility, morphology, and substratum. Our discriminant analysis tools enabled the consideration of many more cell attributes than is customary in cell motility studies. The observations correlate with behaviors seen *in vivo* and suggest specific roles for the extracellular matrix proteins and their integrin receptors in metastasis. Cell translocation *in vitro* has been associated with malignancy, as has an elongated phenotype[88] and a rounded phenotype.[76] Our study suggests that extracellular matrix proteins contribute in different ways to the malignancy of cancer cells, and that multiple malignant phenotypes exist.

Ultrasonic assisted cell disruption for drug delivery.[55] Although biological effects of ultrasound must be avoided for safe diagnostic applications, ultrasound's ability to disrupt cell membranes has attracted interest as a method to facilitate drug and gene delivery. This preliminary study seeks to develop rules for predicting the degree of cell membrane disruption based on specified ultrasound parameters and measured acoustic signals. Too much ultrasound destroys cells, while cell membranes will not open up for absorption of macromolecules when too little ultrasound is applied. The key is to increase cell permeability to allow absorption of macromolecules, and to apply ultrasound transiently to disrupt viable cells so as to enable exogenous material to enter without cell damage. Thus our task is to uncover a "predictive rule" of ultrasound-mediated disruption of red blood cells using acoustic spectrums and measurements of cell permeability recorded in experiments.

Our predictive model and solver for generating prediction rules are applied to data obtained from a sequence of experiments on bovine red blood cells. For each experiment, the attributes consist of four ultrasound parameters, acoustic measurements at 400 frequencies, and a measure of cell membrane disruption. To avoid over-training, various feature combinations of the 404 predictor variables are selected when developing the classification rule. The results indicate that the variable combination consisting of ultrasound exposure time and acoustic signals measured at the driving frequency and its higher harmonics yields the best rule, and our method compares favorably with classification tree and other *ad hoc* approaches, with correct classification rate of 80% upon cross-validation and 85% when classifying new unknown entities. Our methods used for deriving the prediction rules are broadly applicable, and could be used to develop prediction rules in other scenarios involving different cell types or tissues. These rules and the methods used to derive them could be used for real-time feedback about ultrasound's biological effects. For example, it could assist clinicians during a drug delivery process, or could be imported into an implantable device inside the body for automatic drug delivery and monitoring.

Identification of tumor shape and volume in treatment of sarcoma.[53] This project involves the determination of tumor shape for adjuvant brachytherapy treatment of sarcoma, based on catheter images taken after surgery. In this application, the entities are overlapping consecutive triplets of catheter markings, each of which is used for determining the shape of the tumor contour. The triplets are to be classified into one of two groups: Group 1 = triplets for which the middle catheter marking should be bypassed, and Group 2 = triplets for which the middle marking should not be bypassed. To develop and validate a classification rule, we used clinical data collected from 15 soft tissue sarcoma (STS) patients. Cumulatively, this comprised 620 triplets of catheter markings. By careful (and tedious) clinical analysis of the geometry of these triplets, 65 were determined to belong to Group 1, the "bypass" group, and 555 were determined to belong to Group 2, the "do-not-bypass" group.

A set of measurements associated with each triplet is then determined. The choice of what attributes to measure to best distinguish triplets as belonging to Group 1 or Group 2 is non-trivial. The attributes involved distance between each pair of markings, angles, curvature formed by the three triplet markings. Based on the selected attributes, our predictive model was used to develop a classification rule. The resulting rule provides 98% correct classification on cross-validation, and was capable of correctly determining/predicting 95% of the shape of the tumor on new patients' data. We remark that the current clinical procedure requires manual outline based on markers in films of the tumor volume. This study was the first to use automatic construction of tumor shape for sarcoma adjuvant brachytherapy.[53]

Discriminant analysis of biomarkers for prediction of early atherosclerosis.[51] Oxidative stress is an important etiologic factor in the pathogenesis of vascular disease. Oxidative stress results from an imbalance between injurious oxidant and protective antioxidant events in which the former predominate.[66,78] This results in the modification of proteins and DNA, alteration in gene expression, promotion of inflammation, and deterioration in endothelial function in the vessel wall, all processes that ultimately trigger or exacerbate the atherosclerotic process.[17,83] It was hypothesized that novel biomarkers of oxidative stress would predict early atherosclerosis in a relatively healthy nonsmoking population who are free from cardiovascular disease. One hundred and twenty-seven healthy nonsmokers, without known clinical atherosclerosis had carotid intima media thickness (IMT) measured using ultrasound. Plasma oxidative stress was estimated by measuring plasma lipid hydroperoxides using the determination of reactive oxygen metabolites (d-ROMs) test. Clinical measurements include traditional risk factors including age, sex, low density lipoprotein (LDL), high density lipoprotein (HDL), triglycerides, cholesterol, body-mass-index (BMI), hypertension, diabetes mellitus, smoking history, family history of CAD, Framingham risk score, and Hs-CRP.

For this prediction, the patients are first clustered into two groups: (Group 1: $\text{IMT} \geq 0.68$,Group 2: $\text{IMT} < 0.68$). Based on this separator 30 patients belong to Group 1, and 97 belong to Group 2. Through each iteration, the classification method trains and learns from the input training set and returns the most discriminatory patterns among the 14 clinical measurements; ultimately resulting in the development of a prediction rule based on observed values of these discriminatory patterns among the patient data. Using all 127 patients as a training set, the predictive model identified age, sex, BMI, HDLc, Fhx CAD < 60, hs-CRP and d-ROM as discriminatory attributes that together provide unbiased correct classification of 90%, and 93%, respectively, for Group 1 ($\text{IMT} \geq 0.68$) and Group 2 patients ($\text{IMT} < 0.68$). To further test the power of the classification method for correctly predicting the IMT status on new/unseen patients, we randomly selected a smaller patient training set of size 90. The predictive rule from this training set yields 80% and 89% correct rates for predicting the remaining 37 patients into Groups 1 and 2, respectively. The importance of d-ROM as a discriminatory predictor for IMT status was confirmed during the machine learning process, this biomarker was selected in every iteration as the ''machine'' learned and trained to develop a predictive rule to correctly classify patients in the training set. We also performed predictive analysis using Framingham Risk Score and d-ROM; in this case the unbiased correct classification rates (for the 127 individuals) for Groups 1 and 2 are 77% and 84%, respectively. This is the first study to illustrate that this measure of oxidative stress can be effectively used along with traditional risk factors to generate a predictive rule that can potentially serve as an inexpensive clinical diagnostic tool for prediction of early atherosclerosis.

Fingerprinting native and angiogenic microvascular networks through pattern recognition and discriminant analysis of functional perfusion data.[57] The cardiovascular system provides oxygen and nutrients to the entire body. Pathological conditions that impair normal microvascular perfusion can result in tissue ischemia, with potentially serious clinical effects. Conversely, development of new vascular structures fuels the progression of cancer, macular degeneration, and atherosclerosis. Fluorescence-microangiography offers superb imaging of the functional perfusion of new and existent microvasculature, but quantitative analysis of the complex capillary patterns is challenging. We developed an automated pattern-recognition algorithm to systematically analyze the microvascular networks, and then apply our classification model herein to generate a predictive rule. The pattern-recognition algorithm identifies the complex vascular branching patterns, and the predictive rule demonstrates 100% and respectively 91% correct classification on perturbed (diseased) and normal-tissue perfusion. We confirmed that transplantation of normal bone marrow to mice in which genetic deficiency resulted in impaired angiogenesis eliminated predicted differences and restored normal-tissue perfusion patterns (with 100% correctness). The pattern recognition and classification method offers an elegant solution for the automated fingerprinting of microvascular networks that could contribute to better understanding of angiogenic mechanisms and be utilized to diagnose and monitor microvascular deficiencies. Such information would be valuable for early detection and monitoring of functional abnormalities before they produce obvious and lasting effects, which may include improper perfusion of tissue, or support of tumor development.

The algorithm can be used to discriminate between the angiogenic response in a native healthy specimen compared to groups with impairment due to age, or chemical or other genetic deficiency. Similarly, it can be applied to analyze angiogenic responses as a result of various treatments. This will serve two important goals. First, the identification of discriminatory patterns/attributes that distinguish angiogenesis status will lead to a better understanding of the basic mechanisms underlying this process. Because therapeutic control of angiogenesis could influence physiological and pathological processes such as wound and tissue repairing, cancer progression and metastasis, or macular degeneration, the ability to understand it under different conditions will offer new insight in developing novel therapeutic interventions, monitoring and treatment, especially in aging, and heart disease. Thus, our study and the results form the foundation of a valuable diagnostic tool for changes in the functionality of the microvasculature and for discovery of drugs that alter the angiogenic response. The methods can be applied to tumor diagnosis, monitoring and prognosis. In particular, it will be possible to derive microangiographic fingerprints to acquire specific microvascular patterns associated with early stages of tumor development. Such "angioprinting" could become an extremely helpful early diagnostic modality, especially for easily accessible tumors such as skin cancer.

Prediction of protein localization sites. The protein localization database consists of eight groups with a total of 336 instances (143, 77, 52, 35, 20, 5, 2, 2, respectively) with seven attributes.[69] The eight groups are eight localization sites of protein, including cp (cytoplasm), im (inner membrane without signal sequence), pp (perisplasm), imU (inner membrane, uncleavable signal sequence), om (outer membrane), omL (outer membrane lipoprotein), imL (inner membrane lipoprotein), imS (inner membrane, cleavable signal sequence). However, the last four groups are taken out from our classification experiment since the population sizes are too small to ensure significance.

The seven attributes include mcg (McGeoch's method for signal sequence recognition), gvh (von Heijne's method for signal sequence recognition), lip (von Heijne's Signal Peptidase II consensus sequence score), chg (Presence of charge on N-terminus of predicted lipoproteins), aac (score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins), alm1 (score of the ALOM membrane spanning region prediction program), and alm2 (score of ALOM program after excluding putative cleavable signal regions from the sequence).

In the classification we use four groups, 307 instances, with seven attributes. Our classification model selected the discriminatory patterns mcg, gvh, alm1, and alm2 to form the predictive rule with unbiased correct classification rates of 89%, compared to the results of 81% by other classification models.[43]

Pattern recognition in satellite images for determining types of soil. The satellite database consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood.[69] The aim is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number. There are six groups with 4435 samples in the training dataset and 2000 samples in testing dataset; and each sample entity has 36 attributes describing the spectral bands of the image.

The original Landsat Multi-Spectral Scanner image data for this database was generated from data purchased from NASA by the Australian Centre for Remote Sensing. The Landsat satellite data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multi-spectral and radar data, maps indicating topography, land use etc. is expected to assume significant importance with the onset of an era characterized by integrative approaches to remote sensing (for example, NASA's Earth Observing System commencing this decade).

One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is an 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80 m × 80 m. Each image contains 2340×3380 such pixels.

The database is a (tiny) sub-area of a scene, consisting of 82×100 pixels. Each line of data corresponds to a 3 × 3 square neighborhood of pixels completely contained within the 82×100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3×3 neighborhood and a number indicating the classification label of the central pixel. The number is a code for the following six groups: red soil, cotton crop, gray soil, damp gray soil, soil with vegetation stubble, very damp gray soil. Running our classification model, 17 discriminatory attributes were selected to form the classification rule, producing an unbiased prediction with 85% accuracy.

SUMMARY AND CONCLUSION

In the article, we present a class of general-purpose predictive models that we have developed based on the technology of large-scale optimization and support-vector machines.[14,33,49,50,56] Our models seek to maximize the correct classification rate while constraining the number of misclassifications in each group. The models incorporate the following features: (1) the ability to classify any number of distinct groups; (2) allow incorporation of heterogeneous types of attributes as input; (3) a high-dimensional data transformation that eliminates noise and errors in biological data; (4) constraining the misclassification in each group and a reserved-judgment region that provides a safeguard against over-training (which tends to lead to high misclassification rates from the resul ting predictive rule); and (5) successive multi-stage classification capability to handle data points placed in the reserved-judgment region. The performance and predictive power of the classification models is validated through a broad class of biological and medical applications.

Classification models are critical to medical advances as they can be used in genomic, cell, molecular, and system level analyses to assist in early prediction, diagnosis and detection of disease, as well as for intervention and monitoring. As shown in the CpG island study for human cancer, such prediction and diagnosis opens up novel therapeutic sites for early intervention. The ultrasound application illustrates its application to a novel drug delivery mechanism, assisting clinicians during a drug delivery process, or in devising implantable devices into the body for automated drug delivery and monitoring. The lung cancer cell motility offers an understanding of how cancer cells behave under different protein media, thus assisting in the identification of potential gene therapy and target treatment. Prediction of the shape of a cancer tumor bed provides a personalized treatment design, replacing manual estimates by sophisticated computer predictive models. Prediction of early atherosclerosis through inexpensive biomarker measurements and traditional risk factors can serve as a potential clinical diagnostic tool for routine physical and health maintenance, alerting doctors and patients to the need for early intervention to prevent serious vascular disease. Fingerprinting of microvascular networks opens up the possibility for early diagnosis of perturbed systems in the body that may trigger disease (e.g., genetic deficiency, diabetes, aging, obesity, macular degeneracy, tumor formation), identify target sites for treatment, and monitoring prognosis and success of treatment. Determining the type of erythemato-squamous disease and the presence/absence of heart disease helps clinicians to correctly diagnose and effectively treat patients. Thus classification models serve as a basis for predictive medicine where the desire is to diagnose early and provide personalized target intervention. This has the potential to reduce healthcare costs, improve success of treatment, and improve quality-of-life of patients.

A theoretical study will be performed on these models to understand their characteristics and the sensitivity of the predictive patterns to model/parameter variations. The modeling framework for discrete support-vector machines offers great flexibility, enabling one to simultaneously incorporate the features as listed above, as well as many other features. However, deriving the predictive rules for such problems can be computationally demanding, due to the *NP-hard* nature of MIP (Garey and Johnson 1979).[35] We continue to work on improving optimization algorithms utilizing novel cutting plane and branch-and-bound strategies, fast heuristic algorithms, and parallel algorithms.[8,24,45,46–48,58,59]

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation CAREER Young Investigator award.

REFERENCES

[1] Anderson, J. A. Constrained discrimination between k populations. *J. Roy. Stat. Soc. Ser. B* 31:123–139, 1969.

[2] Anderson, T. W. An Introduction to Multivariate Statistical Analysis. 2nd ed. New York: Wiley, 1984.

[3] Bajgier, S. M., and A. V. Hill. An experimental comparison of statistical and linear programming approaches to the discriminant problems. *Decision Sci.* 13:604–618, 1982.

[4] Banks, W. J., and P. L. Abad. An efficient optimal solution algorithm for the classification problem. *Decision Sci.* 22:1008–1023, 1991.

[5] Bennett, K. P., and E. J. Bredensteiner. A parametric optimization method for machine learning. *INFORMS J. Comput.* 9:311–318, 1997.

[6] Bennett, K. P., and O. L. Mangasarian. Multicategory discrimination via linear programming. *Optimiz. Methods Software* 3:27–39, 1993.

[7] Bishop, C. M. Neural Networks for Pattern Recognition. Oxford: Oxford University Press, 1995.

[8] Bixby, R. E., W. Cook, A. Cox, and E. K. Lee. Computational experience with parallel mixed integer programming in a distributed environment. *Ann. Operat. Res. Spl. Issue Parallel Optimiz.* 90:19–43, 1999.

[9] Bradley, P. S., U. M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS J. Comput.* 11:217–238, 1999.

[10] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[11] Brock, G. J., T. H. Huang, C. M. Chen, and K. J. Johnson. A novel technique for the identification of CpG islands exhibiting altered methylation patterns (ICEAMP). *Nucleic Acids. Res.* 29:E123, 2001.

[12] Broffit, J. D., R. H. Randles, and R.V. Hogg. Distribution-free partial discriminant analysis. *J. Am. Stat. Assoc.* 71:934–939, 1976.

[13] Brooks, J. P., A. Wright, C. Zhu, and E. K. Lee. Discriminant analysis of motility and morphology data from human lung carcinoma cells placed on purified extracellular matrix proteins. *Ann. Biomed. Eng.*, 2006 (in review).

[14] Brooks, J. P., and E. K. Lee. Solving a mixed-integer. programming formulation of a multi-category constrained discrimination model. *INFORMS Proc. Artif. Intell. Data Mining*, 2006a (in press).

[15] Brooks, J. P., and E. K. Lee. Mixed integer programming constrained discrimination model for credit screening. In: Proceedings of the Business Industry Symposium, Spring Simulation Multi-conference 2007, 2006b (in press).

[16] Cavalier, T. M., J. P. Ignizio, and A. L. Soyster. Discriminant analysis via mathematical programming: certain problems and their causes. *Comput. Operat. Res.* 16:353–362, 1989.

[17] Chevion, M., E. Berenshtein, and E. R. Stadtman. Human studies related to protein oxidation: protein carbonyl content as a marker of damage. *Free Radic. Res.* 33(Suppl):S99–S108, 2000.

[18] Conway, K. E., B. B. McConnell, C. E. Bowring, C. D. Donald, S. T. Warren, and P. M. Vertino. TMS1, a novel proapoptotic caspase recruitment domain protein, is a target of methylation-induced gene silencing in human breast cancers. *Cancer Res.* 60:6236–6242, 2000.

[19] Costello, J. F., M. C. Fruhwald, D. J. Smiraglia, L. J. Rush, G. P. Robertson, X. Gao, F. A. Wright, J. D.

Feramisco, P. Peltomaki, J. C. Lang, D. E. Schuller, L. Yu, C. D. Bloomfield, M. A. Caligiuri, A. Yates, R. Nishikawa, H. H. Su, N. J. Petrelli, X. Zhang, M. S. O'Dorisio, W. A. Held, W. K. Cavenee, and C. Plass. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* 24:132–138, 2000.

[20]Costello, J. F., C. Plass, and W. K. Cavenee. Aberrant methylation of genes in low-grade astrocytomas. *Brain Tumor Pathol.* 17:49–56, 2000.

[21]Cristianini, N., and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 2000.

[22]Duarte Silva, A. P., and A. Stam. Second order mathematical programming formulations for discriminant analysis. *Eur. J. Operat. Res.* 72:4–22, 1994.

[23]Duda, R. O., P. E. Hart, and D. G. Stork. Pattern Classification, 2nd ed. Wiley, 2001.

[24]Easton, T., K. Hooker, and E. K. Lee. Facets of the independent set polytope. *Math. Program. B* 98:177–199, 2003.

[25]Feltus, F. A., E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. Predicting aberrant CpG island methylation. *Proc. Natl. Acad. Sci.* 100(21):12253–12258, 2003.

[26]Feltus, F. A., E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. DNA signatures associated with CpG island methylation states. *Genomics* 87:572–579, 2006.

[27]Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7:179–188, 1936.

[28]Freed, N., and F. Glover. A linear programming approach to the discriminant problem. *Decision Sci.* 12:68–74, 1981.

[29]Freed, N., and F. Glover. Evaluating alternative linear programming models to solve the two-group discriminant problem. *Decision Sci.* 17:151–162, 1986.

[30]Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. A genomic sequencing protocol that yields a positive display of 5-ethylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* 89:1827–1831, 1992.

[31]Fruhwald, M. C., M. S. O'Dorisio, L. J. Rush, J. L. Reiter, D. J. Smiraglia, G. Wenger, J. F. Costello, P. S. White, R. Krahe, G. M. Brodeur, and C Plass. Gene amplification in NETs/medulloblastomas: mapping of a novel amplified gene within the MYCN amplicon. *J. Med. Genet.* 37:501–509, 2000.

[32]Gallagher, R. J., E. K. Lee, and D. Patterson. An optimization model for constrained discriminant analysis and numerical experiments with iris, thyroid, and heart disease datasets. In: Proceedings of the 1996 American Medical Informatics Association, edited by J. J. Cimino, 1996, pp. 209–213.

[33]Gallagher, R. J., E. K. Lee, and D. A. Patterson. Constrained discriminant analysis via 0/1 mixed integer programming. *Ann. Operat. Res., Spl. Issue on Non-Trad. Approach. Stat. Classif. Regress.* 74:65–88, 1997.

[34]Gardiner-Garden, M., and M. Frommer. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196:261–282, 1987.

[35]Garey, M. R., and D. S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: Freeman, 1979.

[36]Gehrlein, W. V. General mathematical programming formulations for the statistical classification problem. *Operat. Res. Lett.* 5:299–304, 1986.

[37]Gessaman, M. P., and P. H. Gessaman. A comparison of some multivariate discrimination procedures. *J. Am. Stat. Assoc.* 67:468–472, 1972.

[38]Glover, F. Improved linear programming models for discriminant analysis. *Decision Sci.* 21:771–785, 1990.

[39]Glover, F., S. Keene, and B. Duea. A new class of models for the discriminant problem. *Decision Sci.* 19:269–280, 1988.

[40]Gochet, W., A. Stam, V. Srinivasan, and S. Chen. Multigroup discriminant analysis using linear programming. *Operat. Res.* 45:213–225, 1997.

[41]Habbema, J. D. F., J. Hermans, and A. T. Van Der Burgt. Cases of doubt in allocation problems. *Biometrika* 61:313–324, 1974.

[42]Herman, J. G., J. R. Graff, S. Myohanen, B. D. Nelkin, and S. B. Baylin. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. USA* 93:9821–9826, 1996.

[43]Horton, P., and K. Nakai. A probablistic classification system for predicting the cellular localization sites of proteins. *Intell. Systems Mol. Biol.* 4:109–115, 1996.

[44]Koehler, G. J., and S. S. Erenguc. Minimizing misclassifications in linear discriminant analysis. *Decision Sci.* 21:63–85, 1990.

[45]Lee, E. K. Computational Experience with a General Purpose Mixed 0/1 Integer Programming Solver (MIPSOL), Software Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1997.

[46]Lee, E. K. A Linear-programming based parallel cutting plane algorithm for mixed integer programming problems. In: Proceeding for the Third Scandinavian Workshop on Linear Programming, 1999, pp. 22–31.

[47]Lee, E. K. Branch-and-bound methods. In: Handbook of Applied Optimization, edited by M. G. C. Resende and P. M. Pardalos. Oxford University Press, 2001, ISBN 0-19-512594-0.

[48]Lee, E. K. Generating cutting planes for mixed integer programming problems in a parallel distributed memory environment. *INFORMS J. Comput.* 16:1–28, 2004.

[49]Lee, E. K. Discriminant analysis and predictive models in medicine. In: Interdisciplinary Research in Management Science, Finance, and HealthCare, edited by S. J. Deng. Peking University Press, 2006a (to appear).

[50]Lee, E. K. Optimization-based predictive models in medicine and biology. Optimization in Medicine. Computer Science Series. Springer Netherlands, 2006b (to appear).

[51]Lee, E. K., S. Ashfaq, D. P. Jones, S. D. Rhodes, W. S. Weintrau, C. H. Hopper, V. Vaccarino, D. G. Harrison, and A. A. Quyyumi. Prediction of early atherosclerosis in healthy adults via novel markers of oxidative stress and d-ROMs. Working paper, 2007.

[52]Lee, E. K., T. Easton, and K. Gupta. Novel evolutionary models and applications to sequence alignment problems. *Operat. Res. Med. – Comput. Optimiz. Med. Life Sci.* 148:167–187, 2006.

[53]Lee, E. K., A. Y. C. Fung, J. P. Brooks, and M. Zaider. Automated tumor volume contouring in soft-tissue sarcoma adjuvant brachytherapy treatment. *Int. J. Radiat. Oncol. Biol. Phys.* 47(11):1891–1910, 2002.

[54]Lee, E. K., A. Y. C. Fung, and M. Zaider. Automated planning volume contouring in soft-tissue sarcoma adjuvant brachytherapy treatment. *Int. J. Radiat. Oncol. Biol. Phys.* 51:391, 2001: ASTRO 2001.

[55]Lee, E. K., R. Gallagher, A. Campbell, and M. Prausnitz. Prediction of ultrasound-mediated disruption of cell membranes using machine learning techniques and statistical analysis of acoustic spectra. *IEEE Trans. Biomed. Eng.* 51(1):1–9, 2004.

[56] Lee, E. K., R. J. Gallagher, and D. Patterson. A linear programming approach to discriminant analysis with a reserved judgment region. *INFORMS J. Comput*. 15(1):23–41, 2003.

[57] Lee, E. K., S. Jagannathan, C. Johnson, and Z. S. Galis. Fingerprinting native and angiogenic microvascular networks through pattern recognition and discriminant analysis of functional perfusion data.*PNAS* (submitted).

[58] Lee, E. L., and S. Maheshwary. Facets of conflict hypergraphs. *Math. Operat. Res.* (submitted).

[59] Lee, E. K., and J. Mitchell. Computational experience of an interior-point SQP algorithm in a parallel branch-and-bound framework. High Performance Optimization Techniques, edited by H. Frenks, K. Roos, T. Terlaky, S. Zhang. Kluwer Academic Publishers, 1997, pp. 329–347.

[60] Lee, E. K., and T. L. Wu. Classification and disease prediction via mathematical programming. In: Optimization in Medicine, edited by E. Romeijn and P. Pardalos. Kluwer Academic Publishers, 2006 (to appear).

[61] Liittschwager, J. M., and C. Wang. Integer programming solution of a classification problem. *Manage. Sci.* 24:1515–1525, 1978.

[62] Lim, T. S., W. Y. Loh, and Y. S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learn.* 40:203–228, 2000.

[63] Mangasarian, O. L. Mathematical programming in neural networks. *ORSA J. Comput.* 5:349–360, 1993.

[64] Mangasarian, O. L., W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operat. Res.* 43570577.

[65] Mangasarian, O. L. Mathematical programming in data mining. *Data Mining Knowl. Discov.* 1(2):183–201, 1997.

[66] McCord, J. M. The evolution of free radicals and oxidative stress. *Am. J. Med.* 108:652–659, 2000.

[67] McLachlan, G. J. Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley, 1992.

[68] Müller, K. R., S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks* 12(2):181–201, 2001.

[69] Murphy, P. M., and D. W. Aha. UCI Repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, California, 1994.

[70] Nath, R., W. M. Jackson, and T. W. Jones. A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis. *J. Stat. Comput. Simul.* 41:73–93, 1992.

[71] Nemhauser, G. L., and L. A. Wolsey. Integer and Combinatorial Optimization. New York: Wiley, 1988.

[72] Ng, T-H., and R. H. Randles. Distribution-free partial discrimination procedures. *Comput. Math. Appl.* 12A:225–234, 1986.

[73] Patterson, D. A. Three population constrained discrimination. Technical Report, Department of Mathematical Sciences, University of Montana, Missoula, MT, 1996.

[74] Pavur, R., and C Loucopoulos. Examining optimal criterion weights in mixed integer programming approaches to the multiple-group classification problem. *J. Operat. Res. Soc.* 46:626–640, 1995.

[75] Quesenberry, C. P., and M. P. Gessaman. Nonparametric discrimination using tolerance regions. *Ann. Math. Stat.* 39:664–673, 1968.

[76] Raz, A., and A. Ben-Ze'ev. Cell-contact and -architecture of malignant cells and their relationship to metstasis. *Cancer Metastasis Rev.* 6:3–21, 1987.

[77] Rush, L. J., Z. Dai, D. J. Smiraglia, X. Gao, F. A. Wright, M. Fruhwald, J. F. Costello, W. A. Held, L. Yu, R. Krahe, J. E. Kolitz, C. D. Bloomfield, M. A. Caligiuri, and C. Plass. Novel methylation targets in de novo acute myeloid leukemia with prevalence of chromosome 11 loci. *Blood* 97:3226–3233, 2001.

[78] Sies, H. Oxidative Stress: Introductory Comments, edited by H. Sies. Oxidative Stress, pp. 1–8, 1985.

[79] Stam, A. Nontraditional approaches to statistical classification: some perspectives on L_p-norm methods. *Ann. Operat. Res.* 74:1–36, 1997.

[80] Stam, A., and E. A. Joachimsthaler. Solving the classification problem in discriminant analysis via linear and nonlinear programming. *Decision Sci.* 20:285–293, 1989.

[81] Stam, A., and E. A. Joachimsthaler. A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *Eur. J. Operat. Res.* 46:113–122, 1990.

[82] Stam, A., and C. T. Ragsdale. On the classification gap in mathematical-programming-based approaches to the discriminant problem. *Naval Res. Log.* 39:545–559, 1992.

[83] Tahara, S., M. Matsuo, and T. Kaneko. Age-related changes in oxidative damage to lipids and DNA in rat skin. *Mech. Ageing Dev.* 122:415–426, 2001.

[84] Vapnik, V. The Nature of Statistical Learning Theory. Springer-Verlag, 1999.

[85] Wagner, G., P. Tautu, and U. Wolbler. Problems of medical diagnosis – a bibliography. *Methods Inform. Med.* 17:55–74, 1978.

[86] Yan, P. S., C. M. Chen, H. Shi, F. Rahmatpanah, S. H. Wei, C. W. Caldwell, and T. H. Huang. Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res.* 61:8375–8380, 2001.

[87] Yan, P. S., M. R. Perry, D. E. Laux, A. L. Asare, C. W. Caldwell, and T. H. Huang. CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin. Cancer Res.* 6:1432–1438, 2000.

[88] Zimmermann, A., and H. U. Keller. Locomotion of tumor cells as an element of invasion and metastasis. *Biomed. Pharmacoth.* 41:337–344, 1987.

[89] Zopounidis, C., and M. Doumpos. Multicriteria classification and sorting methods: a literature review. *Eur. J. Operat. Res.* 138:229–246, 2002.