Machine Learning for Detection and Diagnosis of Disease

Paul Sajda

Department of Biomedical Engineering, Columbia University, New York, NY 10027;
email: ps629@columbia.edu

Annu. Rev. Biomed. Eng.
2006. 8:8.1–8.29

The *Annual Review of Biomedical Engineering* is online at
bioeng.annualreviews.org

doi: 10.1146/
annurev.bioeng.8.061505.095802

1523-9829/06/0815-
0001$20.00

Key Words

blind source separation, support vector machine, bayesian network, medical imaging, computational biology

Abstract

Machine learning offers a principled approach for developing sophisticated, automatic, and objective algorithms for analysis of high-dimensional and multimodal biomedical data. This review focuses on several advances in the state of the art that have shown promise in improving detection, diagnosis, and therapeutic monitoring of disease. Key in the advancement has been the development of a more in-depth understanding and theoretical analysis of critical issues related to algorithmic construction and learning theory. These include trade-offs for maximizing generalization performance, use of physically realistic constraints, and incorporation of prior knowledge and uncertainty. The review describes recent developments in machine learning, focusing on supervised and unsupervised linear methods and Bayesian inference, which have made significant impacts in the detection and diagnosis of disease in biomedicine. We describe the different methodologies and, for each, provide examples of their application to specific domains in biomedical diagnostics.

INTRODUCTION

Machine learning, a subdiscipline in the field of artificial intelligence (AI), focuses on algorithms capable of learning and/or adapting their structure (e.g., parameters) based on a set of observed data, with adaptation done by optimizing over an objective or cost function. Machine learning and statistical pattern recognition have been the subject of tremendous interest in the biomedical community because they offer promise for improving the sensitivity and/or specificity of detection and diagnosis of disease, while at the same time increasing objectivity of the decision-making process. However, the early promise of these methodologies has resulted in only limited clinical utility, perhaps the most notable of which is the use of such methods for mammographic screening (1, 2). The potential impact of, and need for, machine learning is perhaps greater than ever given the dramatic increase in medical data being collected, new detection, and diagnostic modalities being developed and the complexity of the data types and importance of multimodal analysis. In all of these cases, machine learning can provide new tools for interpreting the high-dimensional and complex datasets with which the clinician is confronted.

Much of the original excitement for the application of machine learning to biomedicine originated from the development of artificial neural networks (ANNs) (e.g., see 3), which were often proclaimed to be "loosely" modeled after computation in the brain. Although in most cases such claims for brain-like computation were largely unjustified, one of the interesting properties of ANNs was that they were shown to be capable of approximating any arbitrary function through the process of learning (also called training) a set of parameters in a connected network of simple nonlinear units. Such an approach mapped well to many problems in medical image and signal analysis and was in contrast to medical expert systems such as Mycin (4) and INTERNIST (5), which, in fact, were very difficult and time consuming to construct and were based on a set of rules and prior knowledge. Problematic with ANNs, however, is the difficulty in understanding how such networks construct the desired function and thus how to interpret the results. Thus, often such methods are used as a "black box," with the ANN producing a mapping from input (e.g., medical data) to output (e.g., diagnosis) but without a clear understanding of the underlying mapping function. This can be particularly problematic in clinical medicine when one must also consider merging the interpretation of the computer system with that of the clinician because, in most cases, computer analysis systems are seen as adjunctive.

As the field of machine learning has matured, greater effort has gone into developing a deeper understanding of the theoretical basis of the various algorithmic approaches. In fact, a major difference between machine learning and statistics is that machine learning is concerned with theoretical issues such as computational complexity, computability, and generalization and is in many respects a marriage of applied mathematics and computer science.

An area in machine learning research receiving considerable attention is the further development and analysis of linear methods for supervised and unsupervised feature extraction and pattern classification. Linear methods are attractive in that their decision strategies are easier to analyze and interpret relative to nonlinear

classification and regression functions, for example, constructed by ANNs. In addition, a linear model can often be shown to be consistent, at least to first order, with underlying physical processes, such as image formation or signal acquisition. Finally, linear methods tend to be computationally efficient, and can be trained online and in real time.

Particularly important for biomedical applications has been the development of methods for explicitly incorporating prior knowledge and uncertainty into the decision-making process. This has led to principled methods based on Bayesian inference, which are well suited for incorporating disparate sources of noisy measurements and uncertain prior knowledge into the diagnostic process.

This review describes recent developments in machine learning, focusing on supervised and unsupervised linear methods and Bayesian inference, which have made significant impact in the detection and diagnosis of disease in biomedicine. We describe the different methodologies and, for each, provide examples of their application to specific domains in biomedical diagnostics.

Biomarkers: anatomic, physiologic, biochemical, or molecular parameters associated with the presence and severity of specific disease states

BLIND SOURCE SEPARATION

Two important roles for machine learning are (*a*) extraction of salient structure in the data that is more informative than the raw data itself (the feature extraction problem) and (*b*) inferring underlying organized class structure (the classification problem). Although strictly speaking the two are not easily separable into distinct problems, we consider the two as such and describe the state of the art of linear methods for both. In this section we focus on unsupervised methods and application of such methods for recovering clinically significant biomarkers.

Linear Mixing

There are many cases in which one is interested in separating, or factorizing, a set of observed data into two or more matrices. Standard methods for such factorization include singular value decomposition (SVD) and principal component analysis (PCA) (6). These methods have been shown to satisfy specific optimality criteria, for example, PCA being optimal in terms of minimum reconstruction error under constraints of orthogonal basis vectors. However, in many cases these criteria are not consistent with the underlying signal/image-formation process and the resultant matrices have little physical relevance. More recently, several groups have developed methods for decomposing a data matrix into two matrices in which the underlying optimality criteria and constraints yield more physically meaningful results (7–14).

Assume a set of observations is the result of a linear combination of latent sources. Such a linear mixing is quite common in signal and image acquisition/formation, at least to a first approximation, and is consistent with underlying physical mixing process, ranging from electroencephalography (15) to acoustics (16). Given $\mathbf{X}$ as a matrix of observations (M rows by N columns) the linear mixing equation is

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

where **A** is the set of mixing coefficients and **S** is a matrix of sources. Depending on the modality, the columns of **X** and **S** are the coordinate system in which the data is represented (i.e., time, space, wavelength, frequency, etc.). The challenge is to recover both **A** and **S** simultaneously given only the observations **X**. This problem is often termed blind source separation (BSS) because the underlying sources are not directly observed and the mixing matrix is not known. BSS methods have been applied to many fundamental problems in signal recovery and deconvolution (17). Most methods that have been developed attempt to learn an unmixing matrix **W**, which when applied to the data **X** yields an estimate of the underlying sources (up to a scaling and permutation),

$$\hat{\mathbf{S}} = \mathbf{WX}. \tag{2}$$

Consider the case when one assumes the rows of **S** (i.e., the source vectors) are random variables that are statistically independent. This implies that the joint distribution of the sources factors,

$$P(s_1, \ldots, s_L) = P(s_1)P(s_2)\ldots P(s_L), \tag{3}$$

where L indicates the number of underlying sources (with each s_i a row in **S**), and $P(.)$ is the probability density function. In most cases L is not known and represents a hyperparameter that must be set or inferred. BSS methods that exploit statistical independence in their optimality criteria are termed independent component analysis (ICA) (see 18 for review). Several approaches have been developed to recover independent sources, the methods distinguished largely by the objective function they employ, e.g., maximum likelihood (19), maximum a posteriori (9), information maximization (20), entropy estimation (21), and mean-field methods (22). In the case of time series, or other types of ordered data, one can also exploit other statistical criteria such as the nonstationarity and utilize simultaneous decorrelation (16, 23–25). Parra & Sajda (15) formulate the problem of BSS as one of solving a generalized eigenvalue problem, where one of the matrices is the covariance matrix of the observations and the other is chosen based on the underlying statistical assumptions on the sources. This view unifies various approaches in simultaneous decorrelation and ICA, together with PCA and supervised methods such as common spatial patterns (CSP) (26).

The attractive property of these decomposition methods is that the recovered components often result in a natural basis for the data, in particular, if one considers some general properties of natural signals. For example, the marginal statistics of many natural signals (or filtered versions of the signals) are highly non-Gaussian (27, 28). Since, by the central limit theorem, linear mixtures of non-Gaussian random variables will result in marginal statistics that are more closely Gaussian, recovering the independent components captures the generative or natural axes of the mixing process.

Nonnegative Matrix Factorization

One particularly useful method for factoring the data matrix **X** under very general and physically realistic constraints is the nonnegative matrix factorization (NMF)

algorithm (7). The basic idea of the NMF algorithm is to construct a gradient descent over an objective function that optimizes **A** and **S**, and, by appropriately choosing gradient stepsizes, to convert an additive update to a multiplicative one. For example, assuming Gaussian noise, one can formulate the problem of recovering **A** and **S** in Equation 1 as a maximum likelihood estimation,

$$\begin{aligned}\mathbf{A}_{ML}, \mathbf{S}_{ML} &= \underset{\mathbf{A}, \mathbf{S}}{\operatorname{argmax}}\, p(\mathbf{X} \mid \mathbf{A}, \mathbf{S}) \\ &= \underset{\mathbf{A}, \mathbf{S}}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|\mathbf{X}-\mathbf{AS}\|^2}{2\sigma^2}} \\ \text{subject to:} \quad & \mathbf{A} \geq 0, \ \mathbf{S} \geq 0, \end{aligned} \tag{4}$$

where σ is the deviation of the Gaussian noise and (**AS**) its mean.

Maximizing the likelihood is equivalent to minimizing the negative log-likelihood, and Equation 4 can be written as,

$$\begin{aligned}\mathbf{A}_{ML}, \mathbf{S}_{ML} &= \underset{\mathbf{A}, \mathbf{S}}{\operatorname{argmin}}(-\log p(\mathbf{X} \mid \mathbf{A}, \mathbf{S})) \\ &= \underset{\mathbf{A}, \mathbf{S}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{AS}\|^2 \\ \text{subject to:} \quad & \mathbf{A} \geq 0, \ \mathbf{S} \geq 0. \end{aligned} \tag{5}$$

One can compute the gradients of the negative log-likelihood function and construct the additive update rules for **A** and **S**,

$$\begin{aligned} A_{i,m} &\leftarrow A_{i,m} + \delta_{i,m}[(\mathbf{XS}^T)_{i,m} - (\mathbf{ASS}^T)_{i,m}] \\ S_{m,\lambda} &\leftarrow S_{m,\lambda} + \eta_{m,\lambda}[(\mathbf{A}^T\mathbf{X})_{m,\lambda} - (\mathbf{A}^T\mathbf{AS})_{m,\lambda}]. \end{aligned} \tag{6}$$

Note that there are two free parameters, which are the step sizes of the updates. Lee & Seung (29) have shown that by appropriately choosing the step sizes, $\delta_{i,m} = \frac{A_{i,m}}{(\mathbf{ASS}^T)_{i,m}}$, $\eta_{m,\lambda} = \frac{S_{m,\lambda}}{(\mathbf{A}^T\mathbf{AS})_{m,\lambda}}$, the additive update rule can be formulated as a multiplicative update rule, with $\mathbf{X} = \mathbf{AS}$ being a fixed point. The multiplicative update rules for **A** and **S** therefore become

$$\begin{aligned} A_{i,m} &\leftarrow A_{i,m} \frac{(\mathbf{XS}^T)_{i,m}}{(\mathbf{ASS}^T)_{i,m}} \\ S_{m,\lambda} &\leftarrow S_{m,\lambda} \frac{(\mathbf{A}^T\mathbf{X})_{m,\lambda}}{(\mathbf{A}^T\mathbf{AS})_{m,\lambda}}, \end{aligned} \tag{7}$$

where convergence of these update rules is guaranteed (29). By formulating the updates as multiplicative rules in Equation 7, we can ensure nonnegative **A** and **S**, given that both are initialized to be nonnegative and the observations, **X**, are nonnegative.

An intuitive understanding of NMF via geometrical considerations can be developed. The manifold of possible solutions specified by the linear mixing equation and nonnegativity constraints represent an M-dimensional polygonal cone spanned by the M rows of **S**. Nonnegativity constraints require that the row vectors of **S**,

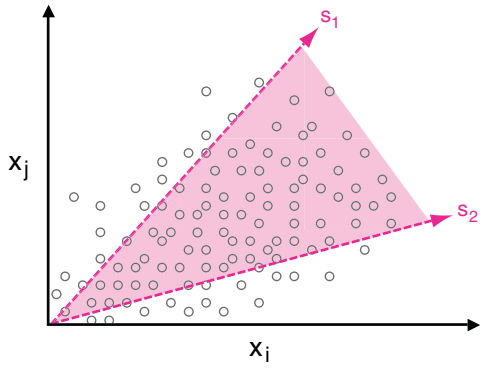


Figure 1

Geometrical interpretation of NMF. The axes represent two dimensions of the high-dimensional space of the observations. Spans of the recovered sources ($\mathbf{s}_1$ and $\mathbf{s}_2$) are shown as dashed magenta vectors. The recovered sources are constrained to lie in the positive hyper-quadrant and tightly envelope the observed data, forming a cone (*pink region*). Points that fall outside of the cone contribute to the error. An analogous picture can be drawn for the basis vectors $\mathbf{A} = \{\mathbf{a}_1 \ldots \mathbf{a}_m\}$.

representing the edges of the cone, lie in the positive quadrant of the *L*-dimensional points defined by the rows of the observations **X**, which must fall within that polygonal cone. The aim of maximum likelihood is to find cone edge vectors that tightly envelope the observed *L*-points. **Figure 1** illustrates this interpretation, which is sometime referred to as a conic encoder (30).

The basic NMF algorithm has been modified in several ways, including adding a sparsity constraint on the sources (31), weighted NMF (32), and constrained NMF (11) (see below). The utility of the NMF algorithm for recovering physically meaningful sources has been demonstrated in a number of application domains, including image classification (33), document classification (34), and separation of audio streams (35), as well as biomedical applications such as analysis of positron emission tomography (PET) (36) and microarray analysis of gene expression (37, 38). Below, we describe two examples, both using nuclear magnetic resonance (NMR) data, where such methods are able to recover signatures of disease and toxicity.

Recovering Spectral Signatures of Brain Cancer

In vivo magnetic resonance spectroscopy imaging (MRSI) allows noninvasive characterization and quantification of molecular markers of potentially high clinical utility for improving detection, identification, and treatment for a variety of diseases, most notably brain cancers (39). MRSI acquires high-frequency resolution MR spectra across a volume of tissue with common nuclei, including ^{1}H (proton), ^{13}C (carbon), ^{19}F (fluorine), and ^{31}P (phosphorus). Machine learning approaches for integrating MRSI with structural MRI have been shown to have potential for improving the assessment of brain tumors (40).

In MRSI, each tissue type can be viewed as having a characteristic spectral profile related to its biochemical composition. In brain tumors, for example, ^{1}H MRSI has shown that metabolites are heterogeneously distributed and, in a given voxel, multiple metabolites and tissue types may be present (41). The observed spectra are therefore a combination of different constituent spectra. Because the signal measured in MRSI is the response to a coherent stimulation of the entire tissue, the amplitudes of different coherent resonators are additive. The overall gain with which a tissue type contributes is proportional to its abundance/concentration in each voxel. As a result, we can explain observations using the linear mixing equation (Equation 1). Because we interpret **A** as abundance/concentration, we can assume the matrix to be nonnegative. In addition, because the constituent spectra **S** represent amplitudes of resonances, in theory, the smallest resonance amplitude is zero, corresponding to the absence of resonance at a given band (where we ignore cases of negative peaks such as in J-modulation). **Figure 2** illustrates the spectral unmixing problem.

Interpretation of MRSI data is challenging, specifically for traditional peak-quantifying techniques (42, 43): A typical dataset consists of hundreds of highly correlated spectra, having low signal-to-noise ratio (SNR) with peaks that are numerous and overlapping. This has created the need for approaches that can analyze the entire dataset simultaneously, taking advantage of the relationships among the spectra to improve the quality of the analysis. Such approaches are particularly useful for spectra

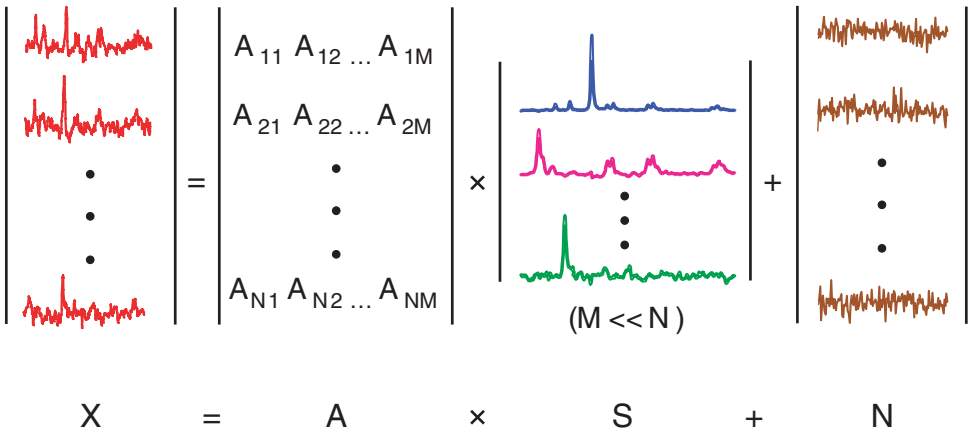


Figure 2

The spectral unmixing problem. Spectra from multiple voxels, for example, from MRSI and represented in the rows of **X**, are simultaneously analyzed and decomposed into constituent spectra **S** and the corresponding intensity distributions **A**. The extracted constituent spectra are identified by comparing them to known spectra of individual molecules. In most cases, the number of rows in **S**, *M*, is much less than the number of rows, *N*, in **X**—i.e., there is a dimensionality reduction in the decomposition. Unidentified spectral components are considered residual noise **N**. Their corresponding magnitudes quantify the modeling error, which can be directly compared to the modeling error of alternative parametric estimation procedures.

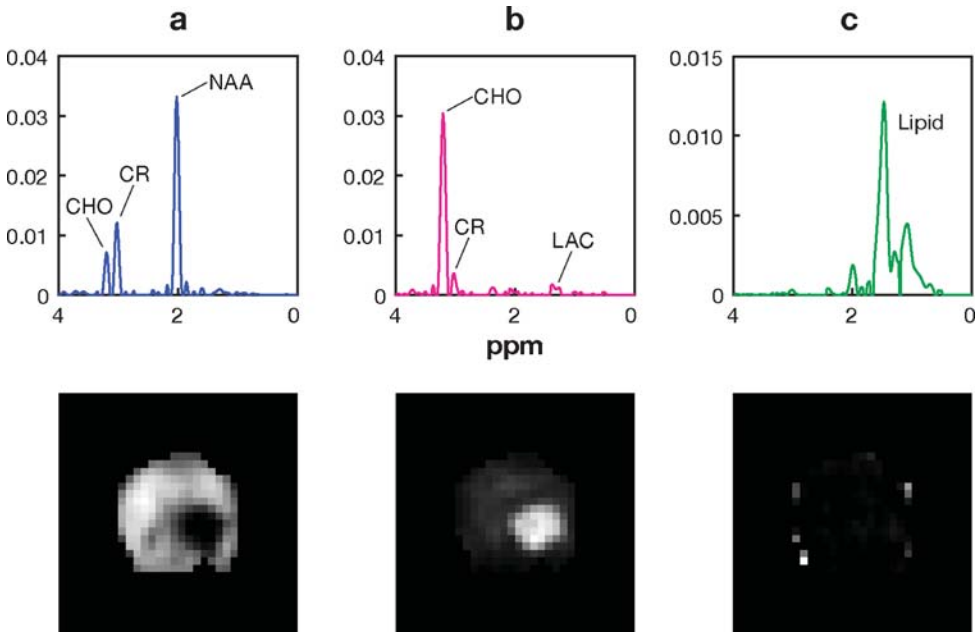


Figure 3

cNMF separation of ^{1}H CSI human brain data into clinically significant biomarkers and their corresponding spatial distributions. (*a*) Spectrum indicative of normal brain tissue: low choline (CHO), high creatine (CR), and high N-acetyl-aspartate (NAA). (*b*) Spectrum indicating high-grade malignant tumor tissue: highly elevated CHO, low CR, almost no NAA, and LAC (lactic acid). (*c*) Spectrum indicating residual lipids.

with low SNR as they utilize the collective power of the data. Several BSS approaches have been developed to simultaneously exploit the statistical structure of an MRSI dataset, factorizing Equation 1. For example, ICA (44), second-order blind identification (SOBI) (45), and bayesian spectral decomposition (8) have all been applied to MRSI datasets to decompose observed spectra into interpretable components.

Constrained NMF (cNMF), is a very efficient version of NMF for recovering biomarkers of brain cancer in MRSI (11, 12). The algorithm enables nonnegative factorization even for noisy observations, which may result in observed spectra having negative values. cNMF includes a positivity constraint, forcing negative values in the recovered spectral sources and abundance/concentration distributions to be approximately zero. **Figure 3** illustrates an example of spectral sources and their corresponding concentrations recovered using cNMF for 1H MRSI data from human brain. In this example, the method recovers biomarkers of high-grade malignant tumor as well as the spatial distribution of their concentration. One of the advantages over other decomposition approaches that have been used in NMR, for example, those based on Monte Carlo sampling (8), is that cNMF is computationally efficient and can be used in near real time, when a patient is in the MR scanner.

Extraction of Metabolic Markers of Toxicity

Metabolomics [sometimes referred to as metabonomics (46)] quantitatively measures the dynamic metabolic response of living systems to pathophysiological stimuli or genetic modification. Metabolomic analysis of biofluids based on high-resolution MRS and chemometric methods are valuable in characterizing the biochemical response to toxicity (47). Interpretation of high-resolution ^{1}H biofluid NMR spectra dataset is challenging, specifically for traditional peak-quantifying techniques: A typical dataset consists of at least tens of highly correlated spectra, with thousands of partially overlapping peaks arising from hundreds of endogenous molecules. This has created the need for approaches that can analyze the entire dataset simultaneously for discriminating between different combinations of metabolites, including their dynamic changes.

PCA is widely used for analyzing metabolomic NMR datasets (48, 49). Although a reasonable approach for preprocessing NMR datasets (50), the PCA decomposition does not lead to physically realizable spectral biomarkers. Physically realistic decompositions are not only useful in terms of visualization, but also in classification of metabolic patterns using machine learning and domain knowledge (51).

Figure 4 illustrates NMF applied to ^{1}H NMR spectra of urine from Han Wistar rats in a hydrazine toxicity experiment. Samples were collected from control rats and those treated with three different doses of hydrazine (75, 90, 120 mg/kg) over a period of 150 h (52). Preprocessing, including normalization of the data, has been described elsewhere (53). The NMF algorithm requires about 300 s (Intel Pentium4 1.2 GHz) to obtain the recovered spectral sources, orders of magnitude faster than other decomposition methods yielding similar results (53). The magnitudes in each dose-group, as a function of time, are shown in **Figure 5*a***, with the identified spectral patterns in **Figure 4*a***. NMF was run 100 times (100 independent initializations), with **Figure 4*a*** showing the mean results (*solid lines*) and variation across runs (*dashed lines*). The small variance demonstrates the robustness and fidelity of the NMF in spectral pattern recovery.

Clear is the association of the four spectral patterns with the hydrazine treatment. In control rats, the first (*filled diamonds*) and second (*filled upper-triangle*) spectral sources maintain almost a constant high level, while the third (*inverted-triangle*) and fourth (*open circle*) are very low. Thus, the first spectral source (Krebs cycle intermediates: citrate and succinate) and second spectral source (2-oxoglutarate) are related to the normal patterns, while the third and fourth (2-aminoadipic acid, taurine and creatine) are related to hydrazine. Indeed, in the treated animals, the normal patterns decrease in response to hydrazine and recover after 36 h, while the other two exhibit reciprocal behaviors during the course of the experiment. The data from the 120 mg/kg dose indicates no sign of recovery at 56 h, at which point the animal was sacrificed.

A visual comparison of the spectral sources recovered using NMF with the first principal components recovered using PCA is shown in **Figure 4*a*,*b***. The PCA components do not represent physically realizable spectra and do not appear to be biomarkers of the metabolic status of the animals. This is further illustrated by

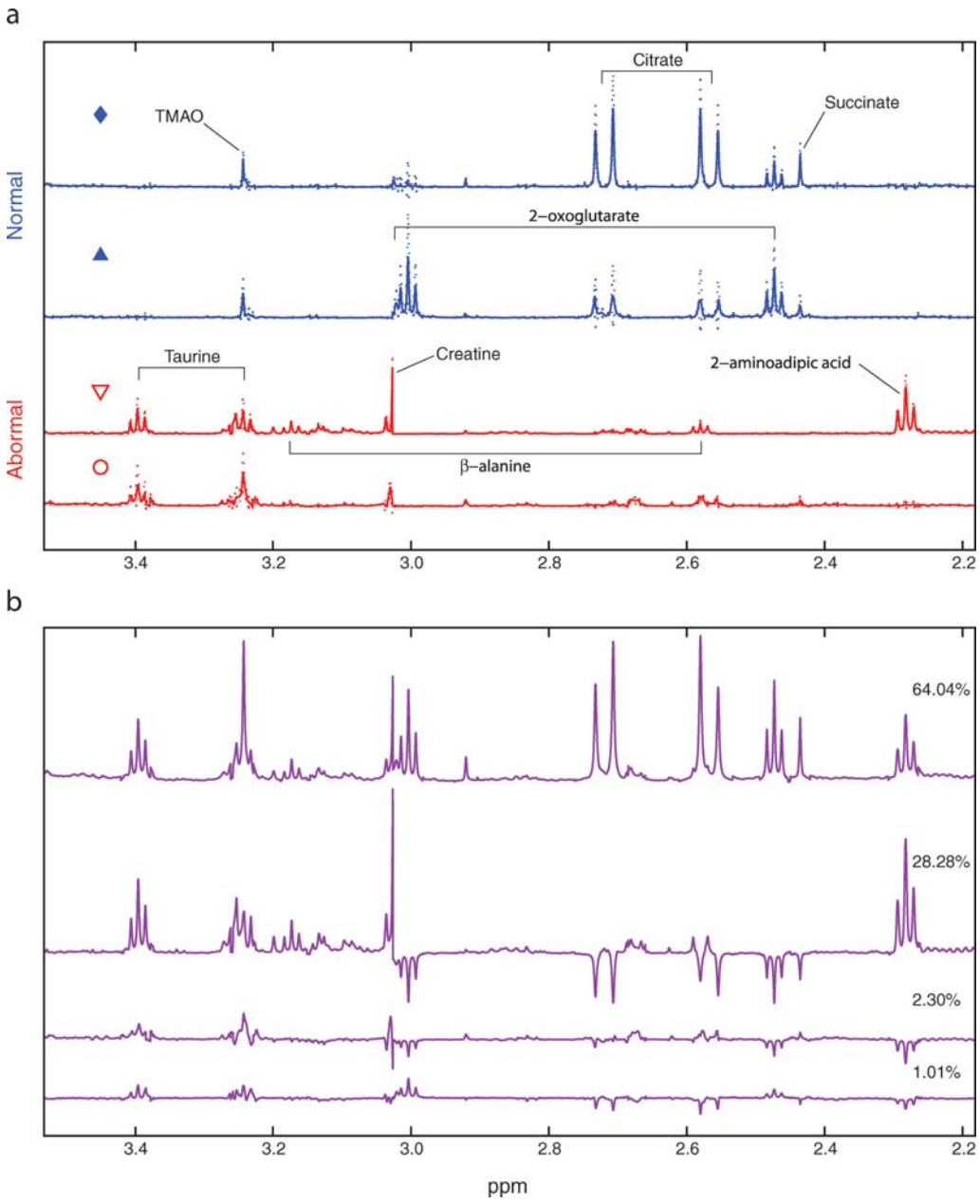


Figure 4

(*a*) Spectral sources, recovered using NMF, indicative of biomarkers for normal metabolic function (*blue*) and hydrazine toxicity (*red*). Solid lines are the mean results and the dash lines are the mean $\pm 2\sigma$. (*b*) Components recovered using PCA. Note that the patterns are not physically realizable spectra because they have negative peaks.

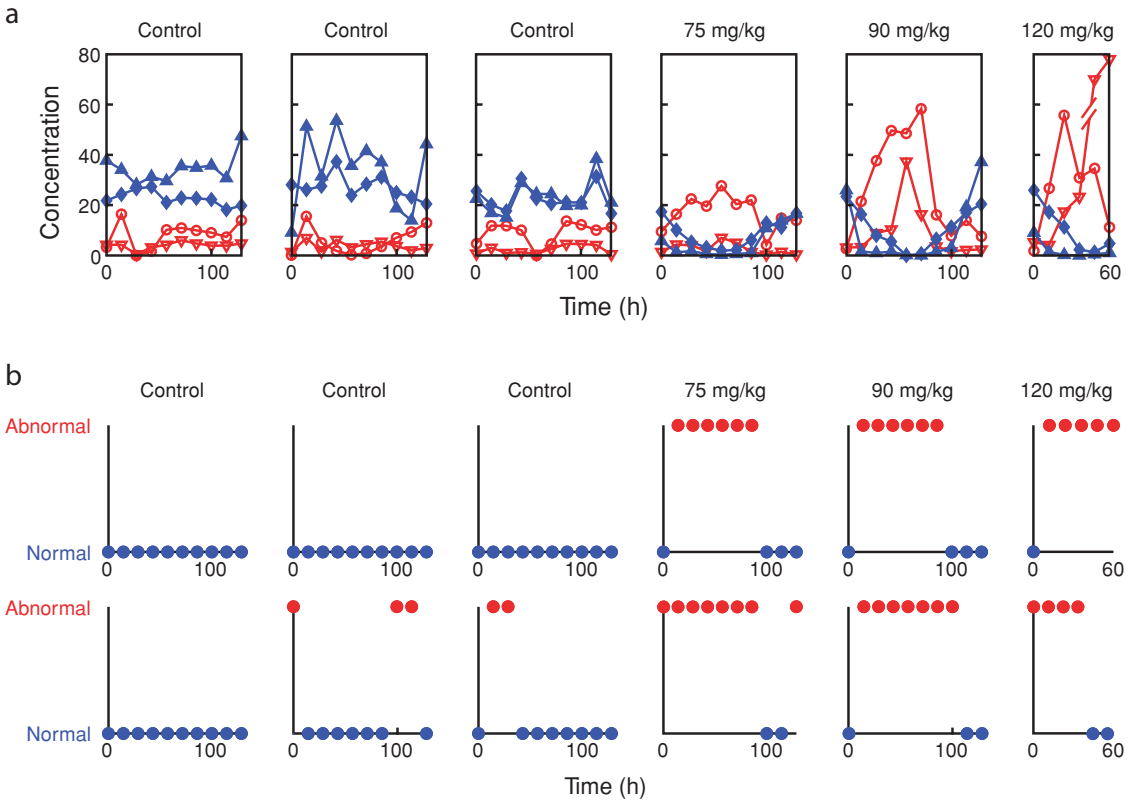


Figure 5

(*a*) Time-dependent concentration of the spectral biomarkers recovered using NMF. The filled diamonds and filled upright triangles are associated with split normal patterns (*blue*), and the inverted triangles and open circles are associated with aberrant patterns (*red*)—symbols correspond to biomarkers in **Figure 4*a***. Analysis of the time-dependent concentrations shows the effect, and in most cases (except the 100 mg/kg dose) recovery from the hydrazine. (*b*) K-means cluster analysis applied to the amplitudes of the NMF patterns and the first four principal components. (*Top*) Concentration profiles recovered via NMF enables correct clustering into normal and abnormal classes. The samples corresponding to the control rats and the ones collected before hydrazine administration, as well as more than 104 h after hydrazine administration for the treated rats, are assigned into the normal cluster, and the other samples collected in the experiment are correctly assigned into the abnormal cluster. (*Bottom*) K-means clustering on the first four principal components. Classification is less accurate compared to when using NMF recovered biomarkers—e.g., as evident by the misclassification of some of the time points for controls.

applying K-means clustering (54) to the amplitudes in the matrix **A** to classify the metabolic status (normal versus abnormal) of the rats as a function of time. The results for clustering the samples into two clusters, normal and abnormal, using cNMF components are shown in **Figure 5*b*** (top), from which we can see that the control rats are clearly separated from those that are treated. Both the initial measurements

(0 h), taken prior to hydrazine administration, and the later data points (after 104 h) for the treated rats are correctly assigned to the normal cluster. These samples have NMR spectra very similar to those from untreated animals, and in fact correspond to time points when the manifested toxic effect of hydrazine is almost minimized by biologic recovery. **Figure 5*b*** (bottom) shows the classification results using the coefficients of the first four PCs. Clearly, these results are less realistic compared with **Figure 5*b*** (top) because some of the time points for the control rats are classified into the abnormal group. We see that a source recovery method that imposes physically realistic constraints improves classification because it connects the recovered sources, quantitatively, with the biological end-point measurements. The approach shows promise for understanding complex metabolic responses of disease, pharmaceuticals, and toxins.

SUPPORT VECTOR MACHINES

The unsupervised learning decompositions discussed in the previous section can be considered methods for constructing descriptive representations of the observed data. An alternative is to construct discriminative representations using supervised learning, namely representations that are constructed to maximize the difference between underlying classes in the data. The most common is linear discrimination. The linear discriminant function can be defined as

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0, \tag{8}$$

and can be seen as defining a hyperplane that maps from the space of the data $\mathbb{D}^n$ to a space of classes $\mathbb{C}^m$, where in most cases $m \ll n$. In binary classification, $m = 1$, and classification is typically done by taking the sign of $f(\mathbf{x})$. An observation $\mathbf{x}$ is mapped into the space of (binary) classes via the weight vector $\mathbf{w}$ and bias w_0. The bias can be absorbed into the weight vector, and in this case it is termed an augmented weight vector (54). The challenge is to learn the weight vector and bias, using supervised methods, which result in minimum classification error, specifically to maximize generalization performance. An illustration of a discriminant function is given in **Figure 6*a***. We can see that there are potentially many ways in which to place a discrimination boundary—i.e., many values for the weights and bias will minimize the classification error. The question thus becomes "Which boundary is the best?" Support vector machines directly address this question.

Hyperplanes and Maximum Margin

A support vector machine (SVM) (see 55–57 for detailed tutorials) is a linear discriminant that separates data into classes using a hyperplane with maximum-margin. Specifically, the discriminant function can be defined using the inner product,

$$f(\mathbf{y}) = \mathbf{w}^{\mathbf{T}}\mathbf{y}, \tag{9}$$

where $\mathbf{y}$ is a result of applying a nonlinear transformation to the data—i.e., $\mathbf{y}_i = \phi(\mathbf{x}_i)$, and classification is done by taking the sign of $f(\mathbf{y})$. The rationale behind the

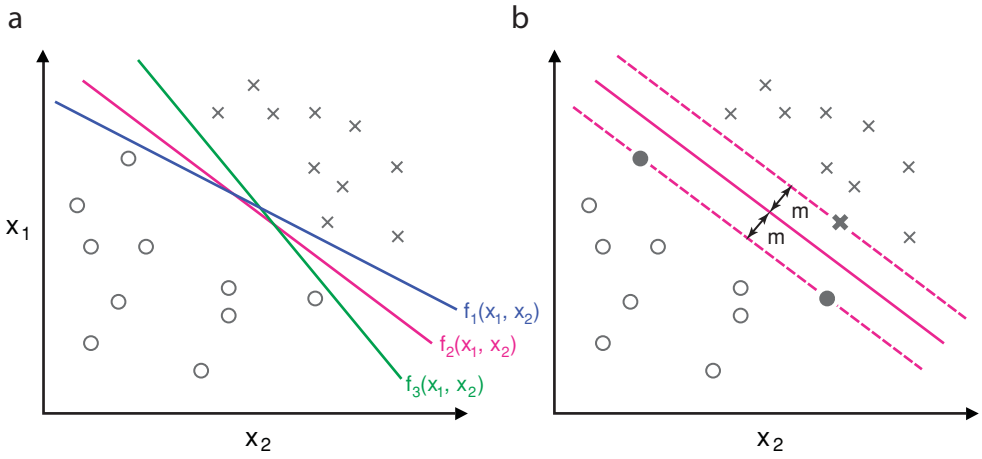


Figure 6

Hyperplanes and maximum margin. (*a*) Two-dimensional scatter plot for a two-class-labeled dataset. The data can be separated by an infinite number of hyperplanes, three of which are shown (f_1, f_2, f_3). (*b*) Illustration of the hyperplane that maximizes the margin (m). This hyperplane is completely specified by the support vectors, those being the example data at the margins.

nonlinear transform is to map the data into a high-dimensional space in which the transformed data is linearly separable and thus divided by a hyperplane. In practice, this transformation is accomplished using the "kernel trick" (58), which enables dot products to be replaced by nonlinear kernel functions—i.e., integral transformation of the function f of the form $(Tf)(y) = \int_a^b k(x, y) f(x) dx$, with the function $k(x, y)$ being the kernel. Much of the current research in the field is focused on developing kernels useful for specific applications and problem domains, where the choice of kernel function embeds some prior knowledge about the problem (e.g., 59–61). The kernel framework is attractive, particularly for applications such as in computational biology, because it can deal with a variety of data types and provide a means for incorporating prior knowledge and unlabeled data into supervised classification.

Bias-variance dilemma: a classic tradeoff encountered in machine learning where one must balance the bias introduced by restricting the complexity of the model with the estimation accuracy or variance of the parameters. The expected generalization error is a combination of the bias and variance and thus the best model simultaneously minimizes these two

For an SVM we learn the hyperplane $\mathbf{w}$ that maximizes the margin between the transformed classes. We can define z_i as an indicator variable which specifies whether a data vector $\mathbf{x}_i$ is in class 1 or 2 (e.g., $z_i = -1$ if $\mathbf{x}_i$ is in class 1 and $z_i = 1$ if $\mathbf{x}_i$ is in class 2). The distance of a hyperplane $\mathbf{w}$ to a (transformed) data vector $\mathbf{y}$ is $|f(\mathbf{y})|/||\mathbf{w}||$. Together with the fact that the separating hyperplane ensures $z_i f(\mathbf{y}_i) \geq 1$ for all n data vectors i, we can express the condition on the margin m as

$$\frac{z_i f(\mathbf{y})}{||\mathbf{w}||} \geq m, i = 1 \ldots n. \quad (10)$$

The goal of SVM training is to find the weight vector $\mathbf{w}$ that maximizes the margin m. Typically, this involves solving a quadratic programming problem. **Figure 6*b*** shows a two-dimensional projection of a separating hyperplane and the corresponding support vectors. Theoretical motivation for SVMs comes from Vapnik Chervonenkis

theory (VC Theory) (62), which provides a test error bound being minimized when the margin is maximized. VC theory can be seen as implementing Occam's Razor.

Occam's (or Ockham's) Razor: principle attributed to the fourteenth-century English logician and Franciscan friar, William of Ockham, which states that the simplest solution that accounts for the data is the best. The principle is important in machine learning because it states that a balance must be maintained between model complexity and error. Closely related to the bias-variance dilemma

Curse of dimensionality: describes the rapid increase in volume of a feature space when the dimensionality of the data is augmented. This is a significant challenge for machine learning because such an increase in volume requires exponentially more examples to adequately sample the space

Closer inspection of **Figure 6*b*** clarifies where SVMs get their name. The training examples nearest to the decision boundary completely determine the boundary and margin. These examples (filled points in **Figure 6*b***) are termed support vectors. They are also sometimes termed proto-types and it is often useful to analyze those examples that are support vectors because one can gain insight into the features of the data that drive the formation of the decision boundary.

As described thus far, the SVM assumes linearly separable data, although perhaps in a transformed space. Cortes & Vapnik (63) considered the case that allowed some of the data to be misclassified and thus did not require linear separability. Such "soft margin" classification finds a hyperplane that splits the training data as best as possible while maximizing the distance to the nearest cleanly split examples.

The support vector method can be extended in several ways. For example, multiclass methods have been developed (64–68) as well as methods for applying the maximum margin approach to regression (62, 69). Support vector regression finds a linear model between the (transformed) input and output, where the output is real valued. This linear model incorporates the idea of a maximum margin by constructing a tube around the linear model that specifies the range at which points can deviate from the model without contributing error—i.e., points lying outside the tube contribute to the error.

SVMs have been applied to a range of biomedical disease detection and diagnosis problems, including detection of oral cancers in optical images (70), polyps in CT colonography (71), and detection of microcalcifications in mammograms (72). A more recent study of several machine learning approaches for microcalcification detection has shown that SVMs yield superior classification performance to a number of other approaches, including ANNs (73).

Analysis of Genetic Microarray Data for Cancer Detection and Diagnosis

Although many machine learning methods have been applied in computational biology and bioinformatics (74), SVMs have received considerable attention (75), specifically for the analysis of gene expression measured via microarrays. Microarrays measure messenger RNA (mRNA) in a sample through the use of probes, which are known affixed strands of DNA. mRNA is fluorescently labeled and those that match the probes will bind. Concentration is measured via the fluorescence. The signals can thus be seen as a set of intensities within a known probe matrix.

One of the challenges using microarray data for classifying tissue types and diagnosing disease is the "curse of dimensionality." The data space is typically high dimensional, with only limited number of examples for training—i.e., the data may have hundreds of dimensions but only tens of examples. For example, Mukherjee et al. (76) used SVMs to classify two types of acute leukemia from microarray samples. Original classification results using self-organizing maps on this data (77) relied on selecting a subset of features (50 of the 7129 genes), based on the training data, to

reduce the dimensionality of the problem. Mukherjee et al. were able to demonstrate better classification performance without the need for feature selection. They used all 7129 genes (the dimensionality of their data) given only 38 training samples and 34 test samples. They also defined confidence intervals for their SVM predictions using a cross-validation technique. These confidence bounds enable them to achieve 100% correct classification of the acute leukemias with 0–4 rejected samples (i.e., samples not classified owing to low confidence).

SVM applications to the classification of colon (78–80) and ovarian (79) cancers in microarray data have also shown promising results. In particular, Furey et al. (79) apply SVMs to multiple types of microarray cancer data (ovarian, colon, and leukemia) and show the approach works well on different datasets and classification problems. Segal et al. (81, 82) use the SVM to classify clear cell carcinoma, which display characteristics of both soft-tissue sarcoma and melanoma. Their classification results, in addition to being highly accurate, provide evidence that clear cell carcinoma is a distinct genomic subtype of melanoma. In addition, SVM analysis, together with hierarchical clustering, uncovers a separate subset of malignant fibrous hystiocytoma. Thus, SVMs can be used to discover new classes and mine the data. A recent study has evaluated various types of classifiers for cancer diagnostics, including SVMs, for classification accuracy using a wide array of gene expression microarray data (83). **Table 1** summarizes these results, which demonstrate the superior performance of SVMs.

Cross-validation: a method typically used in supervised learning where a sample of data is divided into multiple subsets with one subset used to train the algorithm, including selecting features and setting hyperparameters, and the remaining subset(s) used as unbiased testing data to evaluate generalization performance

Table 1 A comparison of multiclass SVM (MC-SVM) and non-SVM approaches for classification results for eight different microarray datasets

	Multicategory classification (%)					Binary classification (%)	
Methods	**BT1**	**BT2**	**L1**	**L2**	**LC**	**PT**	**DLBCL**
MC-SVM							
OVR	**91.67**	77.00	**97.50**	97.32	96.05	**92.00**	**97.50**
OVO	90.56	**77.83**	97.32	95.89	95.59	**92.00**	**97.50**
DAGSVM	90.56	**77.83**	96.07	95.89	95.59	**92.00**	**97.50**
WW	90.56	73.33	**97.50**	95.89	95.55	**92.00**	**97.50**
CS	90.56	72.83	**97.50**	95.89	**96.55**	**92.00**	**97.50**
Non-SVM							
KNN	87.94	68.67	83.57	87.14	89.64	85.09	86.96
NN	84.72	60.33	76.61	91.03	87.80	79.18	89.64
PNN	79.61	62.83	85.00	83.21	85.66	79.18	80.89

Bold indicates the classifier with highest accuracy on the given dataset. BT1, brain tumor dataset 1; BT2, brain tumor dataset 2; L1, leukemia datatset 1; LC, lung cancer; PT, prostate tumor; DLBCL, diffuse large B-cell lymphomas. Multiclass SVMs: OVR, one-versus-rest; OVO, one-versus-one; DAGSVM, directed acyclic graph SVM; WW, method by Weston and Watkins; CS, method by Crammer & Singer. Non-SVMs: KNN, K-nearest neighbor; NN, multi-layer perceptron neural network; PNN, probabilistic neural network. Adapted from Reference 83.

BAYESIAN NETWORKS AND GENERATIVE MODELS

Analysis and classification of biomedical data is challenging because it must be done in the face of uncertainty; datasets are often noisy, incomplete, and prior knowledge may be inconsistent with the measurements. Bayesian decision theory (e.g., see 54) is a principled approach for inferring underlying properties of data in the face of such uncertainty. The Bayesian approach became popular in AI as a method for building expert systems because it explicitly represents the uncertainty in the data and decision-making process. More recently, Bayesian methods have become a cornerstone in machine learning, and in learning theory in general, and have been able to account for a range of inference problems relevant to biological learning (84).

In addition to explicitly dealing with uncertainty, Bayesian approaches can be differentiated from other pattern classification methods by considering the difference between discriminative versus generative models (85, 86). For example, recognition or discriminative probabilistic models estimate $P(C|D)$, the conditional probability of class C given data D. An alternative approach is to construct a generative probabilistic model of the data, which using the aforementioned formulation, would be a model that estimates the class conditional distribution, $P(D|C)$. Such a model has several attractive features for biomedical data analysis. For example, classification is possible by training a distribution for each class and using Bayes'rule to obtain $P(C|D) = P(D|C)P(C)/P(D)$. In addition, novel examples, relative to the training data used to build the model, can be detected by computing the likelihood over each model. The ability to identify novel examples is useful for establishing confidence measures on the output (e.g., should the output of the classifier be "trusted" given that the current data is very different from the training data). In addition, novelty detection can be used to identify new data that might be used to retrain/refine the system. Because essentially any type of data analysis can be formulated given knowledge of the distribution of the data, the generative probabilistic model also can be used to compress (87), suppress noise (88), interpolate, increase resolution, etc. Below, we briefly review Bayesian models that are structured as graphs and consider their application to radiographic image analysis.

Belief Propagation

Solving an inference problem often begins with representing the problem using some form of graphical structure. Examples of such graphical models are Bayesian (or belief) networks and undirected graphs, also known as markov networks (89). In a graphical model, a node represents a random variable and links specify the dependency relationships between these variables (90). The states of the random variables can be hidden in the sense that they are not directly observable, but it is assumed that they have observations related to the state values. Graphical models allow for a compact representation of many classes of inference problems. Once the underlying graphical structure has been constructed, the goal is to infer the states of hidden variables from the available observations. Belief propagation (BP) (91) is an algorithm for solving inference problems based on local message passing. In this section, we focus on

undirected graphical models with pairwise potentials, where it has been shown that most graphical models can be converted into this general form (92).

Let x be a set of hidden variables and y a set of observed Variables, and consider the joint probability distribution of x given y given by

$$P(x_1, \ldots, x_n|y) = c \prod_{i,j} T_{ij}(x_i, x_j) \prod_i E_i(x_i, y_i),$$

where c is a normalizing constant, x_i represents the state of node i, $T_{ij}(x_i, x_j)$ captures the compatibility between neighboring nodes x_i and x_j, and $E_i(x_i, y_i)$ is the local interaction between the hidden and observed variables at location i. In the BP algorithm, this joint probability is approximated by a full factorization in terms of marginal probabilities over x_i:

$$P(x|y) \approx c \prod_i b(x_i).$$

$b(x_i)$ is called the local belief, which is an approximate marginal probability at node x_i.

The belief propagation algorithm iterates a local message computation and belief updates (92). The message $M_{ij}(x_j)$ passed from a hidden node x_i to its neighboring hidden node x_j represents the probability distribution over the state of x_j. In each iteration, messages and beliefs are updated as follows:

$$M_{ij}(x_j) = c \int_{x_i} dx_i \, T_{ij}(x_i, x_j) E_i(x_i, y_i) \prod_{x_k \in N_i/x_j} M_{ki}(x_i)$$

$$b(x_i) = c \, E_i(x_i, y_i) \prod_{x_k \in N_i} M_{ki}(x_i),$$

where N_i/x_j denotes a set of neighboring nodes of x_i except x_j. M_{ij} is computed by combining all messages received by x_i from all neighbors except x_j in the previous iteration and marginalizing over all possible states of x_i (**Figure 7**). The current local belief is estimated by combining all incoming messages and the local observations.

It has been shown that, for singly connected graphs, belief propagation converges to exact marginal probabilities (92). Although how it works for general graphs is

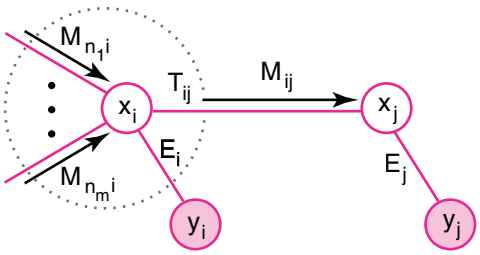


Figure 7

Illustration of local message passing from node x_i to node x_j. Open circles are hidden variables, whereas shaded circles represent observed variables. The local belief at node x_j is computed by combining the incoming messages from all its neighbors and the local interaction E_j.

not well understood, experimental results on some vision problems, such as motion analysis, also show that belief propagation works well for graphs with loops (93). Variants of Bayesian networks include dynamic Bayesian networks (94), useful for constructing generative models of ordered sequential data (e.g., time series). The most well-known type of dynamic Bayesian network is the hidden markov model (95), which has been used, for instance, to model speech. Bayesian networks have been broadly applied in biomedicine, particularly in probabilistic expert systems for clinical diagnosis (96–98) and computational biology (99). They are attractive because they are able to deal with biomedical data that is incomplete or partially correct (100). A novel method for exploiting conditional dependencies in the structure of radiological images to improve detection of breast cancer is described below.

Computer-Aided Diagnosis in Mammography

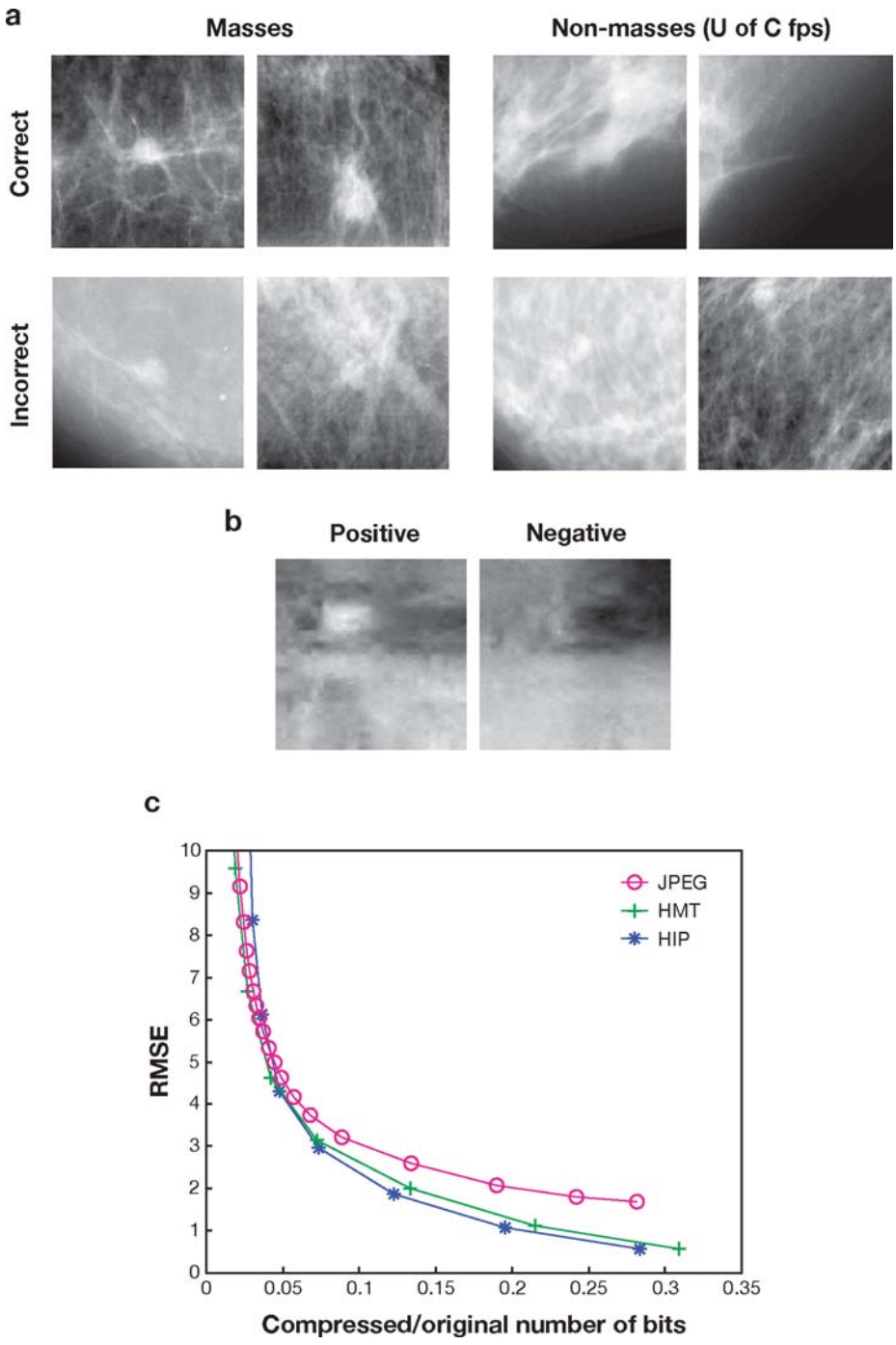
Systems for assisting a radiologist in assessing imagery have been termed computer-aided diagnosis (CAD). CAD is traditionally defined as a diagnosis made by a radiologist who incorporates the results of computer analysis of the imagery (101). The goal of CAD is to improve radiologists' performance by indicating the sites of potential abnormalities, to reduce the number of missed lesions, and/or provide quantitative analysis of specific regions in an image to improve diagnosis.

The shear volume of images collected for screening mammography makes it a prime candidate for CAD. In screening mammography, CAD systems typically operate as automated "second-opinion" or "double-reading" systems that indicate lesion location and/or type. Because individual human observers overlook different findings, it has been shown that double reading (the review of a study by more than one observer) increases the detection rate of breast cancers by 5%–15% (102–104). Double reading, if not done efficiently, can significantly increase the cost of screening, given the need for a second radiologist/mammographer. Methods to provide improved detection with little increase in cost will have significant impact on the benefits of screening. Automated CAD systems are a promising approach for low-cost double reading. Several CAD systems have been developed for mammographic screening and several have been approved by the FDA.

CAD systems for mammography usually consist of two distinct subsystems, one designed to detect microcalcifications and one to directly detect masses (105). A

Figure 8

Generative properties of HIP model for mammographic CAD. (*a*) Example of mammogram regions of interest (ROIs) that the HIP model correctly (*top row*) and incorrectly (*bottom row*) classifies. Note that the difference between the two classes of ROIs (mass versus nonmass) is much more apparent in the top row than in the bottom row, consistent with model performance. (*b*) Mammographic ROI images synthesized by the HIP model. Positive ROIs (*left*) tend to have more focal structure, with more defined borders and higher spatial frequency content. Negative ROI (*right*) tend to be more amorphous with lower spatial frequency content. (*c*) Pixel error (root mean square error, RMSE) versus size of compressed files for JPEG, HIP, and HMT. Clear is that the HIP model results in the best compression. All results (*a–c*) shown for the same HIP model.


a
Masses
Non-masses (U of C fps)
Correct
Incorrect
b
Positive
Negative
c
RMSE
Compressed/original number of bits
JPEG
HMT
HIP
0
1
2
3
4
5
6
7
8
9
10
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35

common element in both subsystems is machine learning algorithms for improving detection and reducing false positive rates introduced by earlier stages of processing. ANNs are particularly popular in CAD because they are able to capture complicated, often nonlinear, relationships in high-dimensional feature spaces not easily captured by heuristic or rule-based algorithms. Several groups have developed neural networks architectures for CAD. Many of these architectures exploit well-known features that might also be used by radiologists (106–109), whereas others utilize more generic feature sets (110–117). In general, these ANNs are discriminative models. Sajda et al. (118) developed a class of generative models for probability distributions of images that are termed hierarchical image probability (HIP) models for application to mammographic CAD. The main elements of the model include the following:

Expectation-maximization (EM) algorithm: an algorithm for finding maximum likelihood estimates of parameters in a probabilistic model where the model depends on both observed and latent (hidden) variables. The algorithm alternates between an expectation (E) step, which computes the expected value of the latent variables, and a maximization step (M), which computes the maximum likelihood estimates of the parameters given the observed variables and the latent variables set to their expectation

- Capturing local dependencies in mammographic images via a coarse-to-fine factoring of the image distribution over scale and position.
- Capturing nonlocal and scale dependencies through a set of discrete hidden variables whose dependency graph is a tree.
- Optimizing model parameters to match the natural image statistics using strict maximum likelihood.
- Enabling both evaluation of the likelihood and sampling from the distribution.
- Modeling the joint distribution of the coefficients of the different subbands at each node as arbitrarily complex distributions using mixtures of Gaussians.
- Separately adjusting the hidden states in each level to better fit the image distribution.
- Using hidden states to capture complex structure in the image through the use of mixture, hierarchy and scale components.

The model exploits the multiscale signatures of disease that are seen in mammographic imagery (119–121) and is trained using the expectation-maximization (EM) algorithm (122), implementing a form of belief propagation. Its structure is similar to other generative models of image distributions constructed on a wavelet tree (123–126).

Figure 8 shows results when training the HIP model on mammographic data. Because the model is generative, a single model can be used for classification, synthesis, and compression. Note, for example, that the synthesis results give some intuition in how the model differentiates masses from nonmass regions of interest (ROIs), namely via focal structure in the image. It is also important to note that with such model of the image distribution we can use the HIP model to achieve better image compression than JPEG or the hidden Markov tree (123).

There are obviously other modalities and medical application areas where generative probabilistic models would be useful. One in particular is multimodal fusion, where the problem is to bring a set of images, acquired using different imaging modalities, into alignment. One method that has demonstrated particularly good performance uses mutual information as an objective criterion (127). The computation of mutual information requires an estimate of entropies, which in turn requires an estimate of the underlying densities of the images. Generative models potentially provide a framework for learning those densities.

CONCLUSION

Machine learning has emerged as a field critical for providing tools and methodologies for analyzing the high volume, high dimensional and multi-modal data generated by the biomedical sciences. This review has provided only a condensed snapshot of applications of machine learning to detection and diagnosis of disease. Fusion of disparate multimodal and multiscale biomedical data continues to be a challenge. For example, current methods have difficulty integrating structural and functional imagery, with genomic, proteomic, and ancillary data to present a more comprehensive picture of disease.

Ultimately, the most powerful and flexible learning machine we know of is the human brain. For this very reason, the machine learning community has become increasingly interested in neuroscience in an attempt to identify new theories and architectures that might account for the remarkable abilities of brain-based learning systems. In fact, Jeff Hawkins, a pioneer in the computer industry, has recently formed a company, Numenta Inc., to begin to develop and productize his theory of how the cortex represents and recognizes patterns and sequences(128). Perhaps, not so coincidentally, early implementation of his theory has been based on hierarchical Bayesian networks, much like those that have been discussed in this review. Thus, the next generation of systems for analyzing biomedical data might ultimately be based on hybrid algorithms that provide the speed and storage of machine systems with the flexibility of human learning.

SUMMARY POINTS

1. Unsupervised matrix decomposition methods, such as nonnegative matrix factorization, which impose general, although physically meaningful, constraints, are able to recover biomarkers of disease and toxicity, generating a natural basis for data visualization and pattern classification.
2. Supervised discriminative models that explicitly address the bias-variance trade-off, such as the support vector machine, have shown great promise for disease diagnosis in computational biology, where data types are disparate and high dimensional.
3. Generative models based on Bayesian networks offer a general approach for biomedical image and signal analysis in that they enable one to directly model the uncertainty and variability inherent to biomedical data as well as provide a framework for an array of analysis, including classification, segmentation, and compression.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation, National Institutes of Health, and the Office of Naval Research Multidisciplinary University Research Initiative.

LITERATURE CITED

1. Nishikawa RM, Haldemann R, Giger M, Wolverton D, Schmidt R, Doi K. 1995. *Performance of a computerized detection scheme for clustered microcalcifications on a clinical mammography workstation for computer-aided diagnosis*. Presented at Radiol. Soc. North Am., p. 425. Chicago, IL (Abstr.)
2. Nishikawa RM, Schmidt RA, Osnis RB, Giger ML, Doi K, Wolverton DE. 1996. Two-year evaluation of a prototype clinical mammographic workstation for computer-aided diagnosis. *Radiology* 201(P):256
3. Bishop CM. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press
4. Shortliffe EH, Buchanan B. 1975. A model of inexact reasoning in medicine. *Math. Biosci.* 23:351–79
5. Miller RA, Pople HE, Myers JD. 1982. Internist-1: an experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.* 307:468–76
6. Jolliffe IT. 1989. *Principal Component Analysis*. New York: Springer-Verlag
7. **Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–91**
8. Ochs MF, Stoyanova RS, Arias-Mendoza F, Brown TR. 1999. A new method for spectral decomposition using a bilinear Bayesian approach. *J. Magn. Reson.* 137:161–76
9. Parra L, Spence C, Ziehe A, Mueller KR, Sajda P. 2000. Unmixing hyperspectral data. In *Advances in Neural Information Processing Systems 12*, ed. SA Solla, TK Leen, K-R Muller, pp. 942–48. Cambrisge, MA: MIT Press
10. Plumbley M. 2002. Conditions for non-negative independent component analysis. *IEEE Signal Proc. Lett.* 9:177–80
11. Sajda P, Du S, Parra L, Stoyanova R, Brown T. 2003. Recovery of constituent spectra in 3D chemical shift imaging using non-negative matrix factorization. *Proc. Int. Symp. Ind. Component Anal. Blind Signal Separation, 4th, Nara, Jpn*, pp. 71–76
12. Sajda P, Du S, Brown TR, Stoyanova R, Shungu DC, et al. 2004. Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *IEEE Trans. Med. Imaging* 23(12):1453–653
13. Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC. 2004. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc. Natl. Acad. Sci. USA* 101(2):641–46
14. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* 100(26):15522–27
15. **Parra L, Sajda P. 2003. Blind source separation via generalized eigenvalue decomposition. *J. Machine Learn. Res. Spec. Iss. ICA* 4:1261–69**
16. Parra L, Spence C. 2000. Convolutive blind source separation of non-stationary sources. *IEEE Trans. Speech Audio Proc.* May:320–27

A much cited paper that describes the nonnegative matrix factorization algorithm and demonstrates its utility for decomposing data into a parts-based structure.

The first to show that many of the common algorithms in independent component analysis could be expressed, together with principal component analysis, as a generalized eigenvalue problem.

17. Sajda P, Zeevi YY, eds. 2005. *Blind Source Separation and Deconvolution in Imaging and Image Processing*, Vol. 15, *Int. J. Imaging Syst. Technol.* New York: Wiley Intersci.
18. Hyvärinen A, Karhunen J, Oja E. 2001. *Independent Component Analysis*. New York: Wiley Intersci.
19. Pearlmutter B, Parra LC. 1995. Maximum likelihood source separation: a context-sensitive generalization of ICA. In *Advances in Neural Information Processing Systems*, ed. MC Mozer, MI Jordan, T Petsche, Vol. 9. Cambridge, MA: MIT Press
20. Bell AJ, Sejnowski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Comp.* 7:1129–59
21. Comon P. 1994. Independent component analysis, a new concept? *Signal Proc.* 36(3):287–314
22. Hojen-Sorensen P, Winther O, Hansen LK. 2002. Mean-field approaches to independent component analysis. *Neural Comp.* 14:889–918
23. Molgedey L, Schuster HG. 1994. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* 72(23):3634–37
24. Cardoso J-F, Souloumiac A. 1993. Blind beamforming for non Gaussian signals. *IEEE Proc. F* 140(6):362–70
25. Belouchrani A, Abed-Meraim K, Cardoso J-F, Moulines E. 1997. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Proc.* 45:434–44
26. Ramoser H, Mueller-Gerking J, Pfurtscheller G. 2000. Optimal spatial filtering of single-trial EEG during imagined hand movements. *IEEE Trans. Rehab. Eng.* 8(4):441–46
27. Wainwright MJ, Simoncelli EP. 1999. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems*, ed. SA Solla, TK Leen, K-R Müller, 12:855–61. Cambridge, MA: MIT Press
28. Parra LC, Spence CD, Sajda P. 2000. Higher-order statistical properties arising from the non-stationarity of natural signals. In *Advances in Neural Information Processing Systems*, pp. 786–92. Cambridge, MA: MIT Press
29. Lee DD, Seung HS. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pp. 556–562. Cambridge, MA: MIT Press
30. Lee DD, Seung HS. 1997. Unsupervised learning by convex and conic coding. In *Advances in Neural Information Processing Systems*, 9:515–21. Cambridge, MA: MIT Press
31. Hoyer PO. 2002. Non-negative sparse coding. In *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland)*, pp. 557–65
32. Guillamet D, Bressan M, Vitria J. 2001. A weighted non-negative matrix factorization for local representations. *IEEE Comput. Soc. Conf. Vision Pattern Recog.*, pages 942–47
33. Guillamet D, Schiele, and J Vitria. Analyzing non-negative matrix factorization for image classification. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume˜2, pages 116–119, 2002.

One of the most cited papers in blind source separation, it introduced the information maximization algorithm (infomax) for recovering sources in instantaneous linear mixtures.

34. W Xu, X Liu, and Y Gong. Document clustering based on non-negative matrix factorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA, 2003. ACM Press.
35. Wang B, Plumbley MD. 2005. Musical audio stream separation by non-negative matrix factorization. *Proc. DMRN Summer Conf.*, Glasgow, Scotland
36. Lee JS, Lee DD, Choi S, Lee DS. 2001. Application of non-negative matrix factorization to dynamic positron emission tomography. *Int. Conf. Ind. Component Anal. Blind Signal Separation, 3rd, San Diego*, ed. T-W Lee, T-P Jung, S Makeig, TJ Sejnowski, pp. 629–32
37. Brunet JP, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* 101(12):4164–69
38. Inamura K, Fujiwara T, Hoshida Y, Isagawa T, Jones MH, et al. 2005. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene* doi:10.1038/sj.onc.1208858
39. Negendank WG, Sauter R, Brown TR, Evelhoch JL, Falini A, et al. 1996. Proton magnetic resonance spectroscopy in patients with glial tumors: a multicenter study. *J. Neurosurg.* 84:449–58
40. Edelenyi FS, Rubin C, Estève F, Grand S, Décorps M, et al. 2000. A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images. *Nat. Med.* 6:1287–89
41. Furuya S, Naruse S, Ide M, Morishita H, Kizu O, et al. 1997. Evaluation of metabolic heterogeneity in brain tumors using ^{1}H-chemical shift imaging method. *NMR Biomed.* 10:25–30
42. Mierisova S, Ala-Korpela M. 2001. MR spectroscopy quantitation: a review of frequency domain methods. *NMR Biomed.* 14(4):247–59
43. Vanhamme L, Huffel SV, Hecke PV, Ormondt DV. 1999. Time domain quantification of series of Biomed. magnetic resonance spectroscopy signals. *J. Magn. Reson.* 140:120–30
44. Ladroue C, Howe FA, Griffiths JR, Tate AR. 2003. Independent component analysis for automated decomposition of in vivo magnetic resonance spectra. *Magn. Reson. Med.* 50:697–703
45. Nuzillard D, Bourg S, Nuzillard J-M. 1998. Model-free analysis of mixtures by NMR using blind source separation. *J. Magn. Reson.* 133:358–63
46. Nicholson JK, Lindon JC, Holmes E. 1999. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica* 29(11):1181–89
47. Keun HC, Ebbels TM, Antti H, Bollard ME, Beckonert O, et al. 2002. Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chem. Res. Toxicol.* 15(11):1380–86
48. Beckwith-Hall BM, Nicholson JK, Nicholls AW, Foxall PJD, Lindon JC, et al. 1998. Nuclear magnetic resonance spectroscopic and principal components

analysis investigations into biochemical effects of three model hepatotoxins. *Chem. Res. Toxicol.* 11(4):260–72
49. Robertson DG, Reily MD, Sigler RE, Wells DF, Paterson DA, Braden TK. 2000. Metabonomics: evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants. *Toxicol. Sci.* 57(2):326–37
50. Stoyanova R, Nicholls AW, Nicholson JK, Lindon JC, Brown TR. 2004. Automatic alignment of individual peaks in large high-resolution spectral data sets. *J. Magn. Reson.* 170(2):329–35
51. Baumgartner C, Böhm C, Baumgartner D. 2005. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *J. Biomed. Inform.* 38(2):89–98
52. Nicholls AW, Holmes E, Lindon JC, Shockcor JP, Farrant RD, et al. 2001. Metabonomic investigations into hydrazine toxicity in the rat. *Chem. Res. Toxicol.* 14(8):975–87
53. Stoyanova R, Nicholson JK, Lindon JC, Brown TR. 2004. Sample classification based on bayesian spectral decomposition of metabonomic NMR data sets. *Anal. Chem.* 76(13):3666–74
54. Duda R, Hart P, Stork D. 2001. *Pattern Classification*. New York: Wiley. 2nd ed.
55. Burges CJC. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov.* 2(2):121–67
56. Cristianni N, Shawe-Taylor J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge Univ. Press
57. Scholkopf B, Burges CJC, Smola AJ. 1998. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press
58. Aizerman M, Braverman E, Rozonoer L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation Remote Contr.* 25:821–37
59. Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. 2001. An introduction to kernel-based learning algorithms 12. *IEEE Neural Networks* 12(2):181–201
60. Scholkopf B, Smola AJ. 2002. *Learning with Kernels*. Cambridge, MA: MIT Press
61. **http://www.kernel machines.org.**
62. Vapnik V. 1999. *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag
63. Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20:273–97
64. Kressel U. 1999. Pairwise classification and support vector machines. In *Advances in Kernel Methods: Support Vector Learning*, Chpt. 15. Cambridge, MA: MIT Press
65. Weston J, Watkins C. 1999. Support vector machines for multi-class pattern recognition. *Proc. Eur. Symp. Artif. Neural Networks (ESANN 99)*, 7th, Bruges
66. Platt JC, Cristianini N, Shawe-Taylor J. 2000. Large margin dags for multiclass classification. In *Advances in Neural Information Processing Systems*, Vol. 12, pp. 547–53. Cambridge, MA: MIT Press

An updated version of the original Duda & Hart (1977), this book is a classic reference in machine learning and pattern classification.

A seminal work in computational learning theory which was the basis for support vector machines.

67. Crammer K, Singer Y. 2000. On the learnability and design of output codes for multiclass problems. *Proc. Annu. Conf. Comp. Learn. Theory (COLT 2000)*, Standford Univ., Palo Alto, CA, June 28–July 1
68. Hsu C-W, Lin C-J. 2002. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks* 13:415–25
69. Scholkopf B, Smola AJ, Williamson RC, Bartlett PL. 2000. New support vector algorithms. *Neural Comp.* 12:1083–121
70. Majumder SK, Ghosh N, Gupta PK. 2005. Support vector machine for optical diagnosis of cancer. *J. Biomed. Optics* 10(2):024034
71. Gokturk SB, Tomasi C, Acar B, Beaulieu CF, Paik DS, et al. 2001. A statistical 3-D pattern processing method for computer-aided detection of polyps in CT colonography. *IEEE Trans. Med. Imaging* 20(12):1251–60
72. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. 2002. A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imaging* 21(12):1552–63
73. Wei L, Yang Y, Nishikawa RM, Jiang Y. 2005. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Trans. Med. Imaging* 24(3):371–80
74. Kapetanovic IM, Rosenfeld S, Izmirlian G. 2004. Overview of commonly used bioinformatics methods and their applications. *Ann, N.Y. Acad. Sci.* 1020:10–21

Provides a thorough review of support vector and other kernel-based machine learning methods applied to computational biology.

75. Scholkopf B, Tsuda K, Vert J-P. 2004. *Kernel Methods in Computational Biology*. Cambridge, MA: MIT Press

76. Mukherjee S, Tamayo P, Slonim D, Verri A, Golub T, et al. 1999. Support vector machine classification of microarray data. Tech. Rep. AI Memo 1677, Mass. Inst. Technol., Cambridge, MA
77. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439):531–37
78. Moler EJ, Chow ML, Mian IS. 2000. Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics* 4:109–26
79. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16(10):906–14
80. Liu Y. 2004. Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci.* 44(6):1936–41
81. Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, et al. 2003. Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am. J. Pathol.* 163(2):691–700
82. Segal NH, Pavlidis P, Noble WS, Antonescu CR, Viale A, et al. 2003. Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. *J. Clin. Oncol.* 21(9):1775–81
83. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21(5):631–43
84. Rao RPN, Olshausen B, Lewicki MS, eds. 2002. *Probabilistic Models of the Brain*. Cambridge, MA: MIT Press

85. Ng AY, Jordan MI. 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, ed. TG Dietterich, S Becker, Z Ghahramani, 14:841-48. Cambridge, MA, MIT Press
86. Jebara T. 2003. *Machine Learning: Discriminative and Generative*. Norwell, MA: Springer
87. Cover TM, Thomas JA. 1991. *Elements of Information Theory*. New York: Wiley
88. Romberg JK, Coi H, Baraniuk RG. 2001. Bayesian tree-stuctured image modeling using wavelet domain hidden markov models. *IEEE Trans. Image Proc.* 10(7):1056–68
89. Smyth P. 1997. Belief networks, hidden markov models, and markov random fields: a unifying view. *Patt. Recogn. Lett.* 18(11–13):1261–68
90. Jordan MI. 2004. Graphical models. *Stat. Sci. (Spec. Iss. Bayesian Stat.)*, 19:140–55
91. **Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann**
92. Yedidia JS, Freeman WT, Weiss Y. 2003. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, ed. G Lakemeyer, B Nebel, pp. 239–69. San Francisco: Morgan Kaufmann
93. Freeman WT, Pasztor EC, Carmichael OT. 2000. Learning low-level vision. *Int. J. Comput. Vision* 40:25–47
94. Ghahramani Z. 1998. Learning dynamic bayesian networks. In *Adaptive Processing of Sequences and Data Structures. Lecture Notes in Artificial Intelligence*, ed. CL Giles, M Gori, pp. 168–97. Berlin: Springer-Verlag
95. Rabiner L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77(2):257–85
96. Andreassen S, Woldbye M, Falck B, Andersen SK. 1987. MUNIN-A causal probabilistic network for interpretation of electromyrographic findings. *Proc. Int. Joint Conf. Artif. Intell.*, 10th, ed. J McDermott, pp. 366–72, Los Altos, CA: Morgan Kaufmann
97. Heckerman DE, Nathwani BN. 1992. An evaluation of the diagnostic accuracy of Pathfinder. *Comput. Biomed. Res.* 25(1):56–74
98. Diez FJ, Mira J, Iturralde E, Zubillaga S. 1997. DIVAVAL, a Bayesian expert system for echocardiography. *Artif. Intell. Med.* 10:59–73
99. **Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805**
100. Nikovski D. 2000. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans. Knowledge Data Eng.* 12(4):509–16
101. Doi K, Giger ML, Nishikawa RM, Hoffmann K, MacMahon H, et al. 1993. Digital radiography: a useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images. *Acta Radiol.* 34:426–39
102. Bird RE. 1990. Professional quality assurance for mammography screening programs. *Radiology* 177:8–10

A major catalyst for research in probabilistic graphical models.

Describes several applications of bayesian network models to gene expression data.

103. Metz CE, Shen JH. 1992. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of roc analysis. *Med. Decision Making* 12:60–75

104. Thurfjell EL, Lernevall KA, Taube AS. 1994. Benefit of independent double reading in a population-based mammography screening program. *Radiology* 191:241–44

105. Giger ML, Huo Z, Kupinski MA, Vyborny CJ. 2000. Computer-aided diagnosis in mammography. In *Handbook of Medical Imaging; Volume 2. Medical Image Processing and Analysis*, ed. M Sonka, JM Fitzpatrick, pp. 917–86. Bellingham, WA: SPIE Press

106. Floyd CE, Lo JY, Yun AJ, Sullivan DC, Kornguth PJ. 1994. Prediction of breast cancer malignancy using an artificial neural network. *Cancer* 74:2944–48

107. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, et al. 1996. Automated feature analysis and classification of malignant and benign microcalcifications. *Radiology* 198:671–78

108. Zheng B, Qian W, Clarke LP. 1996. Digital mammography: mixed feature neural network with spectral entropy decision for detection of microcalcifications. *IEEE Trans. Med. Imaging* 15(5):589–97

109. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. 1998. Automated computerized classification of malignant and benign mass lesions on digital mammograms. *Acad. Radiol.* 5:155–68

An early demonstration of the application of artificial neural networks to computer-assisted diagnosis in mammography. Led to the development of an FDA-approved comprehensive system for mammographic screening.

110. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, Schmidt R. 1994. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Med. Phys.* 21(4):517–24

111. Lo SC, Chan HP, Lin JS, Li H, Freedman MT, Mun SK. 1995. Artificial convolution neural network for medical image pattern recognition. *Neural Networks* 8(7/8):1201–14

112. Zhang W, Doi K, Giger ML, Nishikawa RM, Schmidt RA. 1996. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Med. Phys.* 23:595–601

113. Lo JY, Kim J, Baker JA, Floyd CE. 1996. Computer-aided diagnosis of mammography using an artificial neural network: predicting the invasiveness of breast cancers from image features. In *Medical Imaging 1996: Image Processing*, ed. MH Loew, 2710:725–32. Bellingham, WA: SPIE Press

114. Sajda P, Spence C, Pearson J, Nishikawa R. 1996. Integrating multi-resolution and contextual information for improved microcalcification detection. *Digital Mammography* 96:291–96

115. Chan HP, Sahiner B, Lam KL, Petrick N, Helvie MA, et al. 1998. Computerized analysis of mammographic microcalcifications in morphological and feature spaces. *Med. Phys.* 25:2007–19

116. Sajda P, Spence C. 1998. Applications of multi-resolution neural networks to mammography. In *Advances in Neural Information Processing Systems*, ed. MJ Kearns, SA Solla, DA Cohn, 11:938–44. Cambridge, MA: MIT Press

117. Sajda P, Spence C, Pearson J. 2002. Learning contextual relationships in mammograms using a hierarchical pyramid neural network. *IEEE Trans. Med. Imaging* 21(3):239–50
118. Sajda P, Spence C, Parra L. 2003. A multi-scale probabilistic network model for detection, synthesis and compression in mammographic image analysis. *Med. Image Anal.* 7(2):187–204
119. Li L, Qian W, Clarke LP. 1997. Digital mammography: computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms. *Acad. Radiol.* 11(4):724–31
120. Netsch T, Peitgen HO. 1999. Scale-space signatures for the detection of clustered microcalcifications in digital mammograms. *IEEE Trans. Med. Imaging* 18(9):774–86
121. Sajda P, Laine A, Zeevi Y. 2002. Multi-resolution and wavelet representations for identifying signatures of disease. *Dis. Markers* 18(5–6):339–63
122. Dempster NM, Laird AP, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39:185–97
123. Crouse MS, Nowak RD, Baraniuk RG. 1998. Wavelet-based statistical signal processing using hidden markov models. *IEEE Trans. Signal Proc.* 46(4):886–902
124. Cheng H, Bouman CA. 2001. Multiscale Bayesian segmentation using a trainable context model. *IEEE Trans. Image Proc.* 10(4):511–25
125. Coi H, Baraniuk RG. 2001. Multiscale image segmentation using wavelet-domain hidden markov models. *IEEE Trans. Image Proc.* 10(9):1309–21
126. Wainwright MJ, Simoncelli EP, Willsky AS. 2001. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Appl. Comp. Harmonic Anal.* 11:89–123
127. Wells WM, Viola P, Atsumi H, Nakajima S, Kikinis R. 1996. Multi-modal volume registration by maximization of mutual information. *Med. Image Anal.* 1(1):35–51
128. Hawkins J, Blakeslee S. 2004. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*. Bellingham, WA: Henry Holt

A seminal paper that introduced the expectation-maximization (EM) algorithm for solving maximum likelihood problems with latent variables. The EM algorithm has been broadly used across the machine learning community.