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. 8:537–65

First published online as a
Review in Advance on
April 17, 2006

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

doi: 10.1146/
annurev.bioeng.8.061505.095802

1523-9829/06/0815-
0537$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**,

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.