Integrative modeling for prediction

Alfred O Hero
University of Michigan
Ann Arbor, MI
48109-2122

Abstract

This is a synopsis of the presentation that I gave at the MIT LIDS Paths Ahead symposium in November 2009.

1 Introduction

It is a given that our technological society is becoming increasingly data rich with a literal explosion of diverse data sources. It is inconceivable that any individual could today keep track of the major axes of science and technology in the same way that "les lumières" did during the enlightenment, e.g., think of the polymat Newton, Pascal, and Gauss. Automated discovery methods like data mining have become essential to help us filter and interpret the vast amount of data currently available. Such algorithms will be integrative and be able to separate the relevant from the irrelevant. The field of systems biology is a prime example of this trend, and some stunning advances in large scale integrative analysis have been recently reported, e.g., the Genomic Encyclopedia Project [5].

A basic foundation of scientific method is that any model, formula, or procedure must be testable. To be testable it must be predictive: it must be capable of making educated guesses about outcomes that have not yet been observed. Prediction depends on the underlying question or task, which all but determines the objective function against which the predictive accuracy of different models can be compared. However, it is hard to resist the temptation to dumb-down the objective function in order to make analysis tractable and implementation simple. The intrinsic complexity of the problems typically encountered in signal processing and machine learning make such simplification appealing. But we must resist whenever possible!

There are of course many paths to resist oversimplification that we take. For example, non-parametric or semi-parametric approaches accompanied by empirical risk minimization strategies (active learning, multi-task learning); decomposition of the objective function (modularity, sequential convexity); surrogate objective functions (minimization or maximization of a bound); and performance-driven model-reduction and approximation (PCA, manifold learning). These can be formulated as bottom-up (ensembles of simple models) or top-down (pruning of complex models) approaches and each has advantages in different situations. Regardless of the approach, making an effort to quantitatively and analytically compare performance is essential. However, analysis is extremely challenging in integrative predictive models. Methods must be developed to address these challenges.

This summary is organized as follows. In the next section a motivating application is given that exemplifies the challenges faced by integrative predictive models. This is followed by discussions of the adequacy or inadequacy of present tools and some ideas for future tool development. Finally a short conclusion is given. A list of references cited in the document is included at the end.

2 Motivating example: Predictive Health and Disease

The Wall Street Journal of November 3, 2009, contained an article entitled

> "Is There a Case of the Flu in Your Future? With Pentagon Funding, Scientists Are Devising a Test to Predict Who Will Get Sick Before the Symptoms Set In"

The article reported on a recently DARPA project called Predictive Health and Disease run by Program Manager Col. Geoffrey Ling. According to the WSJ article, the long term objective of the PHD project is "...to develop a portable device, about the size of a BlackBerry, that can quickly determine if someone is on the way to being sick." Since the device must be portable it is not practical for it collect and process massive amounts of information. Thus a major research objective is to determine a small number of highly discriminating biomarkers that could be quickly collected, assayed, and processed by such a device. Experiments have been performed by the research investigators at Duke in order to generate data for analysts at Duke and Michigan. These experiments, called challenge studies, consist of deliberate inoculation of volunteers with a respiratory virus such as H3N2 and H1N1 followed by clinical observation of the subjects over a period of several days. The clinical observations consist of symptoms, blood samples, and samples of other bodily fluids taken at regular intervals. The principal challenge to the analysts is variable selection: to discover what combination of fluids and biological assay, e.g., metabolomic, genomic, and proteomic expression profiles, will provide the best predictive power for predicting who will get sick. One can also consider exogenous variables such as epidemiological information and social interactions of the subjects. Some of the initial findings are described in a recent paper [6] for the H3N2 challenge study.

The amount of data available from these studies greatly exceeds the number of samples (subjects) available for performance assessment and prediction. Thus purely computational models for the predictor are prone to overfitting and generalization error. A mix of computational and biological modeling is required. Biological models are being developed by biology researchers for a number of functional pathways, e.g., innate immune and inflammatory response, but these are not yet organized in such a way that they can be reliably implemented as "soft information," as a constraint or penalty for biomarker discovery. The merging of partial or unreliable information with observed data will be a key to making progress. However, as analysts we know that to do this we require attribution of confidence or accuracy to this information. Unfortunately, today such attributes are seldom available in practice.

3 Adequacy of present tools

There has recently been much work on developing tools for large scale variable selection. Most of these tools rely on the parsimony of intrinsic sparsity: in a sea of variables only a very few of them are relevant. When this is true one can achieve striking gains in computation and performance as compared with standard variable selection such as PCA or exhaustive search. These methods use linear or additive model fitting with soft constraints or penalties to guide the variable selection process. A primary difficulty is that these sparse methods are not yet scalable to the huge state spaces and diverse multi-modal data sets that arise in applications such as predictive health and disease.

Furthermore, the standard assumption of intrinsic sparsity is often inadequate for the predictive health and disease application described above. This is due to the fact that there is a substantial amount of redundancy in biological systems which means that there exist large groups of highly correlated variables. Methods such as logistic group lasso [3] and elastic nets [7] are appropriate for classification-penalized variable selection problems but do not easily accommodate the integration of unreliable biological model information.

These methods also fail when there are “giant components” in the thresholded correlation graph, a situation that occurs frequently in data from biological systems. What is needed is a new generation of methods that can handle these limitations of current methodology.

Statistical analysis and prediction tools are motivated by models for the systematic component, e.g., mean response, and the random components, e.g., noise and other sources of variability away from the mean. All models can be classified into three categories: parametric models, semi-parametric models, and non-parametric models. Each of these categories have well developed procedures for estimation, classification and detection. However, semi-parametric models appear to have received much less attention in signal processing and machine learning communities. The promising recent methods of non-parametric machine learning lie somewhere between these categories.

Virtually all prediction models suffer from the classic tradeoff between overfitting the noise and systematic bias. To cope with this tradeoff a standard approach is to penalize for overfitting with a penalty that measures the intrinsic complexity or degrees of freedom of the candidate model. This approach is an example of decomposition or modularization of the overall objective function, e.g. probability of missclassification error: variable selection is performed by maximizing the sum of two objective functions, one measuring goodness of fit, e.g., hinge loss, and the other measuring complexity of the model, e.g., BIC. However, there is no theory that guarantees that parsimony is modular or decomposable. Hence if variable selection is a primary goal, as in predictive health and disease, it is worth considering other approaches to controlling overfitting error.

The final arbiter is the performance of the predictor and its relative merits and weaknesses with respect to other predictors. In large scale problems involving many variables simulation is not a viable approach to performance prediction. On the other hand, analytical approaches to performance prediction rely on tenuous assumptions. The statisticians toolbox of strong and weak laws of large numbers, central limit theorems, and concentration bounds provide asymptotic expressions for performance that are frequently interpretable and intuitive. However, these expressions provide poor predictions of behavior when the sample size is small relative to the model dimension. Such methods are also frequently only tractable when the observations are i.i.d. and any uncertainty in the assumed model is well characterized. Extensions, or even a new class, of analysis techniques is needed to overcome these deficiencies.

4 Tools of the future?

Crystal ball gazing is always dangerous when one tries to make predictions on scant prospective vision. I will simply list two broad areas of interest to me.

Scalable integrated phenomenological and statistical models: Hybrid continuous/discrete/graph valued models are required for many data integration tasks yet the theory is not yet well understood in the context of scalable variable selection. Embedded simulation approaches such as particle filtering are effective for non-linear models but are not scalable. Combinations of model homogenization (graphical models), function approximation, and stochastic approximation methodologies might be considered. One modest step in this direction is [4] in which PCA and gaussian graphical model approaches were combined for distributed implementation of variable selection.

Integrated information-geometric uncertainty quantification: the natural domain to perform information aggregation is the space of probability models (posterior distributions). When certain pseudo-metrics are uses to measure distance between distributions, e.g., information divergences like Kullback-Liebler, the space of probability models takes on some properties of a metric space, the well known information geometry of Amari. In this space one can perform visualization, estimation and dimensionality reduction jointly in a natural manner that preserves information divergence. Some preliminary advances in this direction are reported in [1] and are applied in [2] to a computer vision problem.

5 Conclusion

Existing tools are often inadequate for dealing with demands of emerging application areas of signal processing and machine learning. One emerging application is predictive health and disease that was discussed in this paper. Some other applications with similar requirements are: tomographic information security, multi modality database indexing and retrieval, and quantum molecular imaging. Such problems will bring challenges and require extensions of existing solution approaches as well as some new ones. Such approaches must be capable of dealing with a massive diversity of data with hopefully low latent dimension. They should fully utilize physical models that are hopefully better characterized than they are today. They should more reliably report on expected performance and provide better guarantees on accuracy and sensitivity. In many cases a human in-the-loop will be required at many more places in the system than simply at the designer and customer endpoints. This will be necessary for adapting the system to changing circumstances such as model variability and evolution of user communities, which may include both cooperative and adversarial components. For such humans-in-the-loop flexible information visualization techniques will be needed that avoid information overload. Finally any viable approach will need to intelligently exploit available domain knowledge, e.g., soft social net information, systems biology models, and behavioral pathways.

References

[1] K. Carter, R. R, W. Finn, and A. Hero, "Fine: Fisher information non-linear embedding," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 31, no. 3, pp. 2093–2098, 2009.

[2] L. Mei, M. Sun, K. Carter, A. Hero III, and S. Savarese, "Unsupervised Object Pose Classification from Short Video Sequences," in *British Machine Vision Conference*, p. `www.eecs.umich.edu/~hero/Preprints/Liang_bmvc_09.pdf`, 2009.

[3] L. Meier, S. van de Geer, and P. Buhlmann, "The group lasso for logistic regression," *J. Royal Statistical Society, Ser. B (Statistical Methodology)*, vol. 70, no. 1, pp. 53, 2008.

[4] A. Wiesel and A. O. Hero, "Decomposable principal components analysis," *IEEE Trans. on Signal Processing*, vol. 57, no. 11, pp. 4369–4378, nov 2009.

[5] D. Wu and etal., "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea," *Nature*, vol. 462, pp. 1056–1060, 24 December 2009.

[6] A. Zaas, M. Chen, J. Varkey, T. Veldman, A. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, et al., "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host & Microbe*, vol. 6, no. 3, pp. 207–217, 2009.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.