Workshop on current challenges in statistical learning (11w5051)

Hugh Chipman (Acadia University) Xiaotong Shen (University of Minnesota)
Robert Tibshirani (Stanford University) Joseph Verducci (Ohio State University)
Ji Zhu (University of Michigan) Mu Zhu (Waterloo University)

December 11, 2011

1 Overview of the Field

In recent years, statistical learning has seen rapid growth within statistics and computer sciences. This growth has been driven primarily by the need to analyze data of complex structures and process massive amounts of data from scientific investigations. In a discovery process, statistical uncertainty is usually high, given the limited amount of information contained in the data. In gene function prediction, for example, data may be structured, and contain features whose size greatly exceeds the sample size. This imposes major challenges to statistical learning, which demands powerful statistical tools to efficiently and accurately extract information of scientific interest from high-dimensional (or massive) data of complex structures.

In statistical learning, challenges arise from real applications, which require the processing of complex combinations of thousands of potential features to make reliable and valid generalizations. As a result, developing efficient and accurate methods has become of paramount importance for large-scale and high-dimensional problems. For instance, support vector machines were designed to optimize estimated margins between distributions for classification rules, bypassing the estimation of distributions for efficiency and accuracy. Emerging issues in applications continue to be a driving force in the development of statistical learning.

Statistical learning analyzes statistical aspects of general types of relations, for data expressed in terms of video sequences, text documents, gene and social networks and web-pages, among others. Areas of interest include unsupervised, semisupervised and supervised learning, rankings, text and web mining, network analysis, genomics, drug discovery, intrusion and fraud detection. In what follows, we list several specific

research areas and directions that have broad interest among potential workshop participants. Kernels Methods and Large Margin Classification: Through the concept of margins, kernel methods [7] classify objects of interest by mapping data onto a feature space, with each coordinate corresponding to one feature of the data. This mapping is efficiently computed through kernels for various types of data. One advantage of kernel methods is that a nonlinear problem is treated linearly after kernel mapping, permitting an efficient treatment of large-scale problems. Kernel methods, known as kernel machines, have been rapidly developing. For kernel methods, several websites have been constructed to communicate recent developments, e.g., www.kernel-machines.org). Despite the extensive successes of kernel-based large margin theory [3], issues remain with regard to its relationship with distribution-based classification theory, as well as how to account for biases in model selection when estimating generalization error. In addition, early attempts at designing a universal kernel have been only modestly successful, and criteria for choosing an appropriate kernel remain a topic of practical interest.

2 Recent Developments and Open Problems

This workshop provided a unique opportunity for researchers to explore the filed of statistical learning in depth; discussed pros and cons of the existing methods. In addition, the workshop identified new research directions from different, but related, areas. Emerging areas for statistical learning include: (a) Ensembles methods for massive high-dimensional data, (b) Regularization, (c) High-dimensional feature selection, and (d) Graphical and network models, and (e) Genomic analysis and data mining.

Ensembles

One of the key developments for ensembles utilizes the notion of combining multiple predictors to form a single accurate predictor. Such ensemble methods take many forms and have resulted in remarkably flexible and efficient tools for prediction. One ensemble method that has generated significant interest is the Boosting [5] family of algorithms such as AdaBoost. Boosting was motivated from the PAC model of learning, which successively puts more weight on misclassified objects until they become correctly classified. These methods have been recently reformulated in the context of additive or logistic models with specialized loss functions. Although the empirical evidence is strong for combining different representations, theoretical understanding of these methods remains lacking. For example, the Winnow algorithm, a kind of perceptron that uses a multiplicative weight-update scheme, performs especially well when many dimensions are irrelevant. In this case it has generalized better than its large margin counterpart. Thus the principles of large margins and low generalization error may, and do, lead to different assessments of classifiers.

Regularization

In statistical learning, regularization [9] introduces additional information to regression; usually related to

the complexity of a solution, which is incorporated as a penalty. For example, in linear regression, a penalty penalizes an increase in a models size through individual regression coefficients; in nonlinear regression a penalty is imposed for lack of smoothness. Regularization with different penalties can lead to solutions to a variety of problems, which is interpretable from the Bayesian perspective. In high-dimensional data analysis, issues continue to arise with respect to how to design suitable penalties and how to tune regularization parameter(s) for predictive accuracy.

High-dimensional Feature Selection

Feature selection is a fundamental tool for data analysis. One focus of recent research has been centered on feature selection in high-dimensional situations, to respond to the pressing need to process large amounts of data of complex structures. In high-dimensional situations, data analysis often involves a large number of features, which may greatly exceed the sample size. In the past, feature selection has been extensively investigated mainly for low-dimensional situations, where many information criteria such as AIC [1] and BIC [6] have been proposed for model selection. For high-dimensional situations, however, developing efficient computational tools becomes extremely important. Recent developments of Lasso [8] and Lars [4] for feature selection have demonstrated the need for efficient computation. Moreover, various methods have been proposed and studied for achieving high predictive accuracy as well as for accuracy of selection. The focus in this area has been on developing efficient computational tools leading to desired statistical properties.

Graphical and network models

Graphical models [2] are useful to analyze and visualize conditional independence relationships between interacting units, in addition to their structural implications. For instance, in dynamic network analysis, a structural change is often a result of certain events or experimental conditions. In Gaussian graphical models, precision matrices are estimated to describe dependencies among interacting units through maximum likelihood. In the past, the research effort has concentrated on reconstruction of a *single* sparse graph. It is known that existing methods may not perform well when the dimension of a matrix is larger than the sample size. For multiple graphical models, detection of structural changes over graphs has been one focus. However, this is challenging due to the enormous size of candidate graphs, which is super-exponential in the total number of nodes over multiple models.

Genomic Applications and Data Mining

Various statistical learning methods have been widely used in genomics analysis, including clustering for microarray analysis, hierarchical classification and semisupervised learning for gene function prediction and gene network analysis, among others. Various emerging issues from biomedical applications are activated in the presence of structured data from various gene networks, where mining the structures of a problem becomes critical.

3 Presentation Highlights

Tensor data analysis. Art Own, Professor of Statistics, at Stanford University, opened the week with an overview of tensor data analysis, where he focused on when and why the bootstrap method breaks for a tensor of three or more. As he indicated, no proper bootstrap can exist in such a situation. He then modified a resampling scheme, which has showed to perform well for the famous Netflix data, which is a sparsely sampled table with rows for customers and columns for movies, or vice versa. His central message is that care is necessary for tensor data analysis. There was also be a case study discussed by Dean Eckles, who presented challenges in analyzing consumer behavior data collected at Facebook, from a practitioner's point of view.

High-dimensional feature selection. High-dimensional feature selection remained to be a focus of this workshop. Wonyul Lee, a graduate student at North Carolina at Chapel Hill, presented results on consistent feature selection in high-dimensional situation. Yongdai Kim, Seoul National University, presented results on feature selection and parameter estimation for nonconvex regularization, and argued that nonconvex regularization with suitably designed regularizers is advantageous over its convex counterpart statistically. Then he discussed the issue of local versus global solutions for nonconvex regularization. Marina Vannucci, Rice University, presented a class of Bayesian models, for feature selection. The models incorporate additional information such as gene functions and gene-gene relations.

Classification and clustering. There were several technical talks on classification and clustering. Yichao Wu, North Carolina State University, proposed weighted learning methods in the context of support vector machines, which aims to solve a nonconvex problem through weighted learning. Ruben Zamar, University of British Columbia, presented a classification method that is robust not only to outliers in the training and also in the test data. This is achieved through an ensemble of robust classifiers based on mixture models. Alejandro Murua presented statistical models behind algorithms for biclustering analysis, which may have nice Bayesian interpretations. Matias Salibian-Barrera, University of British Columbia, presented their results on sparse-K-means clustering algorithms, which have a nice sparseness property, in addition to robustness.

Graphical and network models. Estimation of high-dimensional network structures has become one active area of research recently, which arises naturally in the analyses of many physical, biological and socio-economic systems. Of particular interest is learning the structure of a network over time. George Michailidis, University of Michigan, presented network Granger causal models for exploration of sparsity of its edges and inherent grouping structure among its nodes. George proposed interesting algorithms based on a variant of Group Lasso to discover the Granger causal interactions among the nodes of the network. Of course, there are issues in estimation of relevant covariance and precision matrices. Junhui Wang, University of Illinois at Chicago, argued that positivity of covariance and precision matrices need to be reinforced. He presented

results based on gradient descent algorithms to generate positivity matrices.

Applications. There are several exciting application talks. David van Dyk, Imperial College, London, presented massive data-analytic and data-mining challenges for statistical analysis of astronomic data. Hongzhe Li, University of Pennsylvania, focused on a problem of segment identification. This arises in in studying copy number variants that are alternations of DNA of a genome that results in the cell having a less or more than two copies of segments of the DNA. The problem under consideration is ultra-high dimensional. Li and his collaborator proposed methods for robust identification. In addition, Annie Qu talked about selecting correlation structure for large cluster size data. Identifying correct correlation structure is very important to improve the efficiency of parameter estimation. In their approach, they transformed a correlation structure model selection problem to be a covariate model selection problem, which is capable to handle the increasing cluster size as the sample size increases.

4 Publication of Papers

A poll was taken at the meeting to see if there were substantial interest in publishing papers from the workshop in some sort of proceedings, and a majority were in favor. An on-line survey asked each participant for a level of interest, ranging from not interested (the work has already been submitted elsewhere), to uncertain (need to check with co-authors), to definite intent. With more than half the participants showing some interest, and about 10 stating definite intent, a vote was taken on the exact form of publication. By majority vote it was decided that all interested participants should submit their workshop-related papers for publication in a special issue of Statistical Analysis and Data Mining, a leading journal in this field, which is co-published by the American Statistical Association and Wiley-Blackwell Publishers. One of the co-organizers, Joe Verducci, is Editor-in-Chief of this journal, and he obtained unanimous consent from his Editorial Board in favor of the special issue. A deadline of April 30, 2012 has been set for submission. All papers will undergo the usual review process of the journal, and a special introduction will be written to acknowledge the support from BIRS. Publication is expected in early 2013.

5 Outcome of the Meeting

About 40 people, including statisticians, computer scientists, mathematicians, graduate students, and a broad range of scientists, participate in the week-long workshop. Group discussions, formal and informal, were interleaved with presentations, making for lively exchanges and a creative learning environment.

A very tangible outcome of the meeting, and an indication of the high quality of scientific presentations, was the decision to have a special issue of the journal *Statistical Analysis and Data Mining* devoted to papers presented at the conference. All articles will be subjected to the usual peer-review process for the journal.

Additional details are provided in a separate section below.

Most of the respondents thought that having attended this workshop positively impacted their experience in collaboration and research. Some comments on this impact include:

"I am discussing with Prof. Xiaoming Huo from Georgia Tech a possible project in the near future. He will visit York in March, 2012." *Steven Wang, York University.*

"Yes. Art Own presented an interesting talk on analysis of tensor data, on which I am currently working on." *Junhui Wang University of Illinois at Chicago.*

"[I received an] Invitation to be an AE for Statistical Analysis and Data Mining, which I accepted." *Bertrand Clarke, University of Miami.*

"Stan Young and I talked about a new method of controlling False Discovery Rate that could be used in conjunction with the Tau-Path procedure that I presented. A student of mine is investigating this further." *Joseph Verducci, The Ohio State University.*

"The workshop was very useful to learn about current developments in data mining which are very relevant to my own research. The presentations and discussions with other participants were very inspiring. I also appreciated the immediate feedback I had from other researchers regarding my own current research." *Ruben Zamar, University of British Columbia.*

"Wrote the paper "Robust and Sparse k-means" with Matias Salibian and Yumi Kondo, and Started the new research project "Robust and sparse kernel k-means" with Alejandro Murua, Matias Salibian and Yumi Kondo." *Ruben Zamar, University of British Columbia.*

"I have obtained commitments from several participants at the workshop to submit their latest research to the journal Statistical Analysis and Data Mining (jointly published by Wiley and the American Statistical Association) for which I am currently Editor-in-Chief." *Joseph Verducci, The Ohio State University.*

"I consulted Wenbo Li on a number of results in probability theory. It helped to clarify several technical issues. I am ready to finish up a paper because of this." *Jiahua Chen, University of British Columbia.*

"I've opened a new front in text network models, based on feedback from the meeting; this is joint work with Jacopo Soriano and Justin Gross on models for political blogs." *David Banks, Duke University.*

"David Banks invited me to a SAMSI / SIAM workshop on computational advertising." *Dean Eckles.*

"This workshop has been one of the most positive research experiences I had in the last 5 years! I wish to thank the organizers and the BIRS staff for making this possible." *Ruben Zamar, University of British Columbia.*

6 Wrap-up

The workshop focused on recent developments of machine learning and data mining. Participants of the workshop rated the workshop a success, admitting that many questions have been raised, yet with only a fraction of them have been answered. They all agree that further collaborations are necessary to address a number of emerging issues. First, there are substantial statistical differences between the study of high-dimensional problems and that of conventional problems. Second, efficient algorithms are needed, for performing scale-up data analysis for existing scientific and engineering problems, as well as well those yet to be discovered.

In particular, the workshop identifies several emerging research areas which would be most challenging yet extremely important since they are fundamental problems in machine learning and data mining. One important area is on optimization for high-dimensional non-convex problems. It is urgent to develop feasible algorithm to obtain the global solution. Most of existing theoretical properties are established based on the knowledge that the global solution can be achieved. However, in practice, it is an extremely challenging problem to find the global solution for high-dimensional data. Another important area is to extract important signals from very noise data through matrix decomposition. This is equivalent to obtain low rank approximation from a very high-dimensional matrix. This research area has many applications such as image process, data compression and storage for extremely large data sets in genomics and astronomy studies. The third important area is on high-dimensional network data which are applicable for social network and gene network. Developing fast and efficient algorithm for network data is very challenging since there are typically no replicates available but the number of parameters involved could be very high. Finally, it is also very important to develop a dynamic model for network data in order to evaluate the dynamic changes of network associations over time.

References

[1] Akaike, H. (1973). Information theory and the maximum likelihood principle. In *International Symposium on Information Theory*. Edited by Petrov, V. and Csáki, F. Akademiai Kiádo, Budapest, pp 267-281.

[2] Buhlmann, P., and Van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

[3] Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 73-297.

[4] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32**, 407-499.

[5] Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and Sys. Sci.*, **55**, 119–139.

[6] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-64.

[7] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

[8] TIBSHIRANI, R. J. (1996). Variable selection via Regression shrinkage and selection via LASSO. *J. Royal Statist. Assoc., Ser. B*, **58**, 267-288.

[9] Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet. Math. Dokl.* **5**, 1035-1038.