Statistics and Its Interface Volume 2 (2009) 255–256

Editorial: On the interface of statistics and machine learning

Andreas Christmann and Xiaotong Shen

This special issue in *Statistics and Its Interface* is devoted to several statistical topics in machine learning which obtain great interdisciplinary research interests. In statistics, machine learning has been also extensively studied under the name of nonparametric classification and regression. In the past few years, statistics has seen very significant developments particularly in the areas of nonparametric classification and regression. These developments are driven by a fundamental understanding of the phenomenon of the bias-variance trade-off in function estimation and many applications in science and engineering. While computer scientists and engineers deal with decision functions for classification, statisticians focus more on conditional probabilities that yield decision functions.

These developments are mainly driven by the following facts. There is still an increasing interest in real applications to automatically analyze large and high-dimensional data sets with unknown complex dependency structures basically without any prior knowledge on the data generating distribution. We mention here only data-, text-, and web-mining, fraud detection, non-parametric classification and regression. The computational resources have become available to store huge data sets on the hard-disk and to process these data sets on faster computers, perhaps even using parallel CPUs. Many advances were made in the fundamental understanding of the learning process.

There exists a vast body of literature on statistical learning, especially for classification and regression purposes, and no attempt will be made to give an overview. We would like to mention here only

- *general learning* [9, 12],
- *ensemble methods* [10, 11, 19, 4, 5],
- *neural networks* [18, 2, 17, 1],
- *spline methods* [27, 14, 28],
- *support vector machines* [24, 25, 7, 20, 22],
- *trees* [6, 17, 13], and
- *wavelets* [8, 15].

Ensemble methods are based on the idea to generate many so-called “weak” decision rules and aggregate their results. Special cases are *boosting* [11], *bagging* [4], and *random forests* [5]. In boosting for classification purposes, successive trees give extra weight to points incorrectly predicted by earlier decision rules and finally a weighted vote is taken for the actual prediction. In bagging, successive trees do not depend on earlier trees because each tree is independently constructed from a bootstrap sample of the data set and finally a simple majority vote is taken for the actual prediction. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Important questions of ensemble methods are the choice of weak decision rules and the stopping rule such that the resulting learning method yields an accurate prediction rule for new unseen data, can be computed in an efficient manner, and has good statistical properties.

Support vector machines (SVMs) are nowadays investigated as special minimizers of regularized empirical risks over reproducing kernel Hilbert spaces, although this is (for the special case of the least squares loss function) a relatively old idea (see, e.g., [16, 27] and the references therein). The current view on SVMs with general loss functions is often preferable to the historically used view on SVMs that was based on a geometric idea which led to the first algorithms named “support vector machines,” see [3] and even the *generalized portrait algorithm* proposed by [26]. Large margin classification, on one hand, can be cast into the framework of empirical risk regularization [21]. Under rather weak assumptions SVMs exist, are unique, and depend continuously on the data points. Important questions of SVMs are the choice of the loss function and the kernel (which specifies the reproducing kernel Hilbert space) such that the SVM can “learn” the unknown distribution, has good finite-sample and asymptotic properties, and allows an efficient numerical computation even for large data sets.

The aim of this special issue is to examine important issues in machine learning and statistics to explore ideas for accurate learning. The issue is structured around topics on support vector machines, boosting, regression, and random forests, as well as applications in biomedical research.

With regard to large margin classification, *Zhu, Pan and Shen* propose support vector machines for gene-network classification involving high-dimensional microarray data. *Park and Liu* study the connection between unbounded loss functions and bounded loss functions with respect to accuracy of classification. *Christmann, Van Messem and Steinwart* study consistency and robustness aspects of SVMs which work for all distributions, even for heavy-tailed ones.

With regard to boosting, *Steinwart* investigates generalization properties of boosting classifiers, which is in a parallel fashion as those of large margin classifiers. *Zhu* proposes a multi-class version of two-class AdaBoost. *Kim, Kim, Kim and Lee* embed boosting into the functional ANOVA analysis. *Zhang and Wang* study on another ensemble method–random forest, where they show that a certain random forest can be small enough to achieve a high level of prediction while remaining visible for interpretation and presentation.

Many learning tasks are often high-dimensional. *Hwang, Zhang and Ghosal* combine the idea of selection with shrinkage for high-dimensional sparse regression. *Breheny and Huang* propose feature selection methods through regularization.

Finally, we sincerely hope that this special issue can stimulate further interest of statisticians, computer scientists and engineers, and promote further interdisciplinary collaborations among them to attack important science and engineering problems.

Received 27 June 2009

REFERENCES

[1] Anthony, M. and Bartlett, P. L. (1999).*Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge. MR1741038

[2] Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford. MR1385195

[3] Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM, Madison, WI.

[4] Breiman, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140. MR1425957

[5] Breiman, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

[6] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, CA. MR0726392

[7] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

[8] Daubechies, I. (1991). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.

[9] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York. MR1383093

[10] Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285. MR1348530

[11] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Sys. Sci* **55** 119–130. MR1473055

[12] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

[13] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York. MR1851606

[14] He, X. and Ng, P. (1999). Quantile splines with several covariates. *J. Statist. Plann. Inference* **75** 343–352. MR1678981

[15] Heil, C. and Walnut, D. F. (eds.) (2006). *Fundamental Papers in Wavelet Theory*. Princeton University Press, Princeton, New Jersey. MR2229251

[16] Poggio, T. and Girosi, F. (1990). A theory of networks for approximation and learning. *Proc. IEEE* **78** 1481–1497.

[17] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge. MR1438788

[18] Rosenblatt, R. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Washington, DC. MR0135635

[19] Schapire, R. (1990). The strength of weak learnability. *Mach. Learn.* **5** 197–227.

[20] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA. MR1949972

[21] Smola, A., Bartlett, P. L., Schölkopf, B., and Schuurmans, D. (2000). *Advances in large margin classifiers*. The MIT press, Cambridge, MA. MR1820960

[22] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

[23] Vapnik, V. N. (1982). *Estimation of Dependencies Based on Empirical Data*. Springer, New York. MR0672244

[24] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. MR1367965

[25] Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York. MR1641250

[26] Vapnik, V. N. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Autom. Remote Control* **24** 774–780.

[27] Wahba, G. (1990). *Spline Models for Observational Data*. Series in Applied Mathematics 59, SIAM, Philadelphia. MR1045442

[28] Zhou, S. and Shen, X. (2001). Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *J. Amer. Statist. Assoc.* **96** 247–259. MR1952735

Andreas Christmann
University of Bayreuth
Department of Mathematics
D-95440 Bayreuth
Germany
E-mail address: andreas.christmann@uni-bayreuth.de

Xiaotong Shen
University of Minnesota
School of Statistics
Minneapolis, MN 55455
USA
E-mail address: xshen@stat.umn.edu