ICPR 2006 TUTORIAL

Principles of Stochastic Discrimination and Ensemble Learning

Tin Kam Ho
Bell Labs, Lucent Technologies
tkh@research.bell-labs.com

1 Abstract

Learning in everyday life is often accomplished by making many random guesses and synthesizing the feedback. Kleinberg's analysis of this process resulted in a new method for classifier design – stochastic discrimination (SD). The method constructs an accurate classifier by combining a large number of very weak discriminators that are generated essentially at random. An important advantage is that classifiers designed by this way are insensitive to overtraining.

SD is an ensemble learning method in an extreme form. Studies on other ensemble methods for classification have long suffered from the difficulty of modeling the complementary strengths of the components. The SD theory addresses this rigorously via mathematical notions of enrichment, uniformity, and projectability.

In this tutorial we explain these concepts via a simple numerical example, with a focus on a fundamental symmetry in point set covering that is the key observation leading to the foundation of the SD theory. We illustrate, step by step, how the SD principle operates in this example. We then describe and discuss more sophisticated implementations for practical uses. We believe a basic understanding of the SD method will open the way to explorations of a new classifier technology, and lead to developments of better tools for analyzing other ensemble methods.

2 Detailed Descriptions

Methods for classifier combination, or ensemble learning, can be divided into two categories: 1) *decision optimization* methods that try to obtain *consensus* among a *given* set of classifiers to make the best decision; 2) *coverage optimization* methods that try to *create* a set of classifiers that work best with a *fixed* decision combination function to cover all possible cases.

Decision optimization methods rely on the assumption that the given set of classifiers, typically of a small size, contain sufficient expert knowledge about the application domain, and each of them excels in a subset of all possible input. A decision combination function is chosen or trained to exploit the individual strengths while avoiding their weaknesses. Popular combination functions include majority/plurality votes, sum/product rules, rank/confidence score combination, and probabilistic methods.

While numerous successful applications of these methods have been reported, the joint capability of the classifiers sets an intrinsic limitation on decision optimization that the combination functions cannot overcome. A challenge in this approach is to find out the "blind spots" of the ensemble and to obtain a classifier that covers them.

Coverage optimization methods use an automatic and systematic mechanism to generate new classifiers with the hope of covering all possible cases. A fixed, typically simple, function is used for decision combination. This can take the form of training set subsampling, such as stacking, bagging, and boosting, feature subspace projection, superclass/subclass decomposition, or other forms of random perturbation of the classifier training procedures. Open questions in these methods are 1) how many classifiers are enough? 2) what kind of differences among the component classifiers yields the best combined accuracy? 3) how much limitation is set by the form of the component classifiers?

Apparently both categories of ensemble methods run into some dilemma. Should the component classifiers be weakened in order to achieve a stronger whole? Should some accuracy be sacrificed for the known samples to obtain better generalization for the unseen cases? Do we seek agreement, or differences among the component classifiers?

A central difficulty in studying the performance of these ensembles is how to model the complementary strengths among the classifiers. Many proofs rely on an assumption of statistical independence of component classifiers' decisions. But rarely is there any attempt to match this assumption with observations of the decisions. Often, global estimates of the component classifiers' accuracies are used in their selection, while in an ensemble what matter more are the local estimates, plus the relationship between the local accuracy estimates on samples that are close neighbors in the feature space.

Deeper investigation of these issues leads back to three major concerns in choosing classifiers: discriminative power, use of complementary information, and generalization power. A complete theory on ensembles must address these three issues simultaneously. Many current theories rely, either explicitly or implicitly, on ideal assumptions on one or two of these issues, or have them omitted entirely, and are therefore incomplete.

Kleinberg's theory and method of stochastic discrimination (SD) is the first attempt to explicitly address these issues simultaneously from a mathematical point of view. In this theory, rigorous notions are made for discriminative power, complementary information, and generalization power of an ensemble. A fundamental symmetry is observed between the probability of a fixed model covering a point in a given set and the probability of a fixed point being covered by a model in a given ensemble. The theory establishes that, these three conditions are sufficient for an ensemble to converge, with increases in its size, to the most accurate classifier for the application.

Kleinberg's analysis uses a set-theoretic abstraction to remove all the algorithmic details of classifiers, features, and training procedures. It considers only the classifiers' decision regions in the form of point sets, called weak models, in the feature space. A collection of classifiers is thus just a sample from the power set of the feature space. If the sample satisfies a uniformity condition, i.e., if its coverage is unbiased for any local region of the feature space, then a symmetry is observed between two probabilities (w.r.t. the feature space and w.r.t. the power set, respectively) of the same event that a point of a particular class is covered by a component of the sample. Discrimination between classes is achieved by requiring some minimum difference in each component's inclusion of points of different classes, which is trivial to satisfy. By way of this symmetry, it is shown that if the sample of weak models is large, the discriminant function, defined on the coverage of the models on a single point and the class-specific differences within each model, converges to poles distinct by class with diminishing variance.

We believe that this symmetry is the key to the discussions on classifier combination. However, since the theory was developed from a fresh, original, and independent perspective on the problem of learning, there have not been many direct links made to the existing theories. As the concepts are new, the claims are high, the published algorithms appear simple, and the details of more sophisticated implementations are not known, the method has been poorly understood and is sometimes referred to as mysterious.

It is the goal of this tutorial to illustrate the basic concepts in this theory and remove the apparent mystery. We present the principles of stochastic discrimination with a very simple numerical example. The example is so chosen that all computations can be easily traced step-by-step by hand or with very simple programs. We use Kleinberg's notation wherever possible to make it easier for the interested readers to follow up on the full theory in the original papers. The emphasis in the example is on explaining the concepts of uniformity and enrichment, and the behavior of the discriminant when both conditions are fulfilled.

We then describe a practical implementation of the SD method, and through example applications, illustrate how direct control can be put on the three most important factors affecting classification accuracy – discriminative power and mutual independence of component discriminators, and their ability to generalize. The SD theory also led to some derived methods including random decision forests which place algorithmic control on the three factors in different ways. We describe how such methods relate to the underlying theory.

Informally, the SD method captures the intuition of gaining wisdom from graded random guesses. Kleinberg's mathematical analysis characterizes the essential conditions for such a learning approach to succeed. We discuss broader implications of this analysis on ensemble learning. For example, in the framework of SD theory, boosting can be viewed as a uniformity promoting mechanism. Recent successes of some genetic algorithms on classification (e.g. the XCS classifier) can also be analyzed with the SD theory.

References

[1] E.M. Kleinberg, web site for overview, background, papers, and other information related to SD, `http://kappa.math.buffalo.edu/sd/`.

Main papers on SD theory and method

[2] E.M. Kleinberg, Stochastic discrimination, *Annals of Mathematics and Artificial Intelligence*, **1**, 1990, 207-239, `http://kappa.math.buffalo.edu/sd/sd12r.pdf`.

[3] E.M. Kleinberg, An overtraining-resistant stochastic modeling method for pattern recognition, *Annals of Statistics*, **4**, 6, Dec 1996, 2319-2349, `http://kappa.math.buffalo.edu/sd/aos.pdf`.

[4] E.M. Kleinberg, On the algorithmic implementation of stochastic discrimination, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-22**, 5, May 2000, 473-490, `http://kappa.math.buffalo.edu/sd/473.pdf`.

[5] E.M. Kleinberg, A mathematically rigorous foundation for supervised learning, in J. Kittler, F. Roli, (eds.), *Multiple Classifier Systems*, Lecture Notes in Computer Science 1857, Springer, 2000, 67-76. `http://kappa.math.buffalo.edu/sd/mrfsl2.pdf`.

Related experiments, reviews, tutorials, talks

[6] E.M. Kleinberg, T.K. Ho, Pattern Recognition by Stochastic Modeling, *Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition*, Buffalo, NY, May 1993, 175-183, `http://cm.bell-labs.com/who/tkh/papers/sd1.pdf`.

[7] T.K. Ho, E.M. Kleinberg, Building projectable classifiers of arbitrary complexity, *Proceedings of the 13th International Conference on Pattern Recognition*, Vienna, Austria, August 25-30, 1996, 880-885. `http://cm.bell-labs.com/who/tkh/papers/sd2.pdf`.

[8] A. Shekhawat, SD results for NIPS 2000 competition (top rank for problems 5 and 9), http://q.cis.uoguelph.ca/∼skremer/Research/NIPS2000/.

[9] T.K. Ho, Multiple classifier combination: Lessons and next steps, in A. Kandel, H. Bunke, (eds.), *Hybrid Methods in Pattern Recognition*, World Scientific, 2002. http://cm.bell-labs.com/who/tkh/papers/lessons.pdf.

[10] T.K. Ho, Classifier generating methods and stochastic discrimination, Invited lecture, "International School on Neural Nets Eduardo R. Caianiello" on Ensemble Methods for Learning Machines, IIASS-Vietri sul Mare (Salerno)-ITALY, September 22-28, 2002. Manuscript to appear, slides and notes available at: http://www.cs.bell-labs.com/who/tkh/talks.html.

[11] D.C. Chen, P. Huang, X.Z. Cheng, A concrete statistical realization of Kleinberg's stochastic discrimination for pattern recognition. part I. Two Class classification, *Annals of Statistics*, **31**, 5, Oct 2003, 1393-1412.

Derived works on theory

[12] R. Berlind, *An Alternative Method of Stochastic Discrimination with Applications to Pattern Recognition*, Doctoral Dissertation, Department of Mathematics, State University of New York at Buffalo, 1994.

[13] D. Chen, *Estimates of Classification Accuracies for Kleinberg's Method of Stochastic Discrimination in Pattern Recognition*, Doctoral Dissertation, Department of Mathematics, State University of New York at Bufaflo, 1998.

Derived works on methods

[14] T.K. Ho, Random decision forests, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, August 14-18, 1995, 278-282. http://cm.bell-labs.com/who/tkh/papers/odt.pdf.

[15] T.K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 8, August 1998, 832-844. http://cm.bell-labs.com/who/tkh/papers/df.pdf.

[16] T.K. Ho, Nearest neighbors in random subspaces, *Proceedings of the Second International Workshop on Statistical Techniques in Pattern Recognition*, Sydney, Australia, August 11-13, 1998, 640-648. http://cm.bell-labs.com/who/tkh/papers/rnn.pdf.

[17] D.C. Lambert, software implementation of SD, 2002, http://www.panix.com/∼dcl/.