Diss. ETH No. 16979

Extensions of L_2Boosting and an Application of LogitBoost

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Sciences

presented by
ROMAN WERNER LUTZ
Dipl. Math. ETH
born July 1, 1976
citizen of Zurich

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Sara van de Geer, co-examiner

2007

Abstract

One goal of statistical learning is to establish a relationship between input variables and an output variable. The simplest and most widely used method is linear regression. Over the years a lot of sophisticated methods have been developed. One of them is boosting, which is an ensemble method. This means that it combines several outputs of a simple method to a strong "committee". This PhD thesis proposes various extensions of boosting.

A first extension uses a conjugate direction method instead of the gradient method to construct a new boosting algorithm (CDBoost). As a result, one obtains a fast forward stepwise variable selection algorithm. The conjugate direction of CDBoost is analogous to the constrained gradient in boosting. Using this analogy, CDBoost is generalized to: (i) include small step sizes (shrinkage) which often improves prediction accuracy; (ii) the non-parametric setting with fitting methods such as trees or splines, where least angle regression and the Lasso seem to be unfeasible. The step size in CDBoost has a tendency to govern the degree between L_0- and L_1-penalisation. This makes CDBoost surprisingly flexible. The different methods are compared on simulated and real datasets. CDBoost achieves the best predictions mainly in complicated settings with correlated covariates, where it is difficult to determine the contribution of a given covariate to the response. The gain of CDBoost over boosting is especially high in sparse cases with high signal-to-noise ratio and few effective covariates.

A second extension proposes multivariate L_2Boosting based on some squared error loss for multivariate data. It can be applied to multivariate linear regression with continuous responses and to vector au-

toregressive time series. It is proven, for i.i.d. data as well as for time series, that multivariate L_2Boosting can consistently recover sparse high-dimensional multivariate linear functions, even when the number of predictor variables $p = p_n$ and the dimension of the response $q = q_n$ grow almost exponentially with sample size n, i.e. $p_n = q_n = O(\exp(Cn^{1-\xi}))$ $(0 < \xi < 1, 0 < C < \infty)$, but the L_1-norm of the true underlying function is finite. This theory seems to be among the first to address the issue of large dimension of the response variable; the relevance of such settings is briefly outlined. Some cases where the multivariate L_2Boosting is better than multiple application of univariate methods to single response components are also identified, thus demonstrating that the multivariate approach can be very useful.

Five robustifications of L_2Boosting for linear regression with various robustness properties are considered in the third extension. The first two use the Huber loss as implementing loss function for boosting and the second two use robust simple linear regression for the fitting in L_2Boosting (i.e. robust base learners). Both concepts can be applied with or without down-weighting of leverage points. The last method uses robust correlation estimates and appears to be most robust. Crucial advantages of all methods are that they don't compute covariance matrices of all covariates and that they don't have to identify multivariate leverage points. When there are no outliers, the robust methods are only slightly worse than L_2Boosting. In the contaminated case though, the robust methods outperform L_2Boosting by a large margin. Some of the robustifications are also computationally highly efficient and therefore well suited for high dimensional problems.

Finally, I applied LogitBoost with a tree-based learner to the five performance prediction challenge datasets of the world congress on computational intelligence (WCCI) 2006. The number of iterations and the tree size were estimated by 10-fold cross-validation. A simple shrinkage strategy was added to make the algorithm more stable. The results are promising: I won the challenge.

Zusammenfassung

Ein Ziel von statistischem Lernen ist es, Zusammenhänge zwischen Input Variablen und einer Output Variablen zu finden. Die einfachste und am meisten verwendete Methode ist lineare Regression. Im Laufe der Zeit wurden viele raffinierte Methoden entwickelt. Eine davon ist Boosting, welche eine "Ensemble" Methode ist. Das heisst, sie vereint mehrere Outputs einer einfachen Methode zu einer starken Gesamtheit. Diese Dissertation beschreibt verschiedene Erweiterungen von Boosting.

Eine erste Erweiterung verwendet die Methode der "konjugierten Richtungen" anstatt der Gradientenmethode um einen neuen Boosting Algorithmus zu konstruieren (CDBoost). Als Resultat erhält man einen schnellen Algorithmus für Variablenwahl vorwärts. Die konjugierte Richtung von CDBoost ist analog zum eingeschränkten Gradienten beim Boosting. Unter Verwendung dieser Analogie wird CDBoost so verallgemeinert, dass man (i) kleine Schrittlängen ("Shrinkage") verwenden kann, was sehr oft die Genauigkeit von Vorhersagen verbessert und dass man (ii) auch nicht-parametrische Methoden wie Bäume oder Splines verwenden kann, was bei "least angle regression" und dem Lasso nicht möglich zu sein scheint. Die Schrittlänge in CDBoost hat eine Tendenz, zwischen der L_0- und der L_1-Bestrafung zu regeln. Das macht CDBoost überraschend flexibel. Die verschiedenen Methoden werden anhand von simulierten und echten Datensätzen verglichen. CDBoost erreicht die besten Vorhersagen vorallem in komplizierten Situationen mit korrelierten Kovariablen, in denen es schwierig zu bestimmen ist, wieviel eine Kovariable zur Zielgrösse beiträgt. Der Gewinn von CDBoost über Boosting ist besonders gross in Fällen mit hohem Signal-zu-Rauschen Verhältnis und wenigen effektiven Kovariablen.

Die zweite Erweiterung schlägt multivariates L_2Boosting vor, welches auf einer quadratischen Verlustfunktion für multivariate Daten basiert. Es kann bei multivariater linearer regression mit kontinuierlichen Zielgrössen und für vektorielle autoregressive Zeitreihen verwendet werden. Wir beweisen für i.i.d. Daten als auch für Zeitreihen, dass multivariates L_2Boosting dünn besetzte, hoch-dimensionale, multivariate, lineare Funktionen konsistent schätzen kann und zwar sogar dann, wenn die Anzahl der erklärenden Variablen $p = p_n$ und die Dimension der Zielgrösse $q = q_n$ fast exponentiell wachsen mit der Anzahl Beobachtungen n, d.h. $p_n = q_n = O(\exp(Cn^{1-\xi}))$ $(0 < \xi < 1, 0 < C < \infty)$, aber die L_1-Norm der wahren zugrunde liegenden Funktion endlich ist. Diese Theorie scheint unter den ersten zu sein, die die Situation der hoch-dimensionalen Zielvariable behandeln. Die Relevanz solcher Situationen wird kurz umrissen. Es werden auch Fälle aufgezeigt, in denen das multivariate Boosting besser ist als mehrere Anwendungen von univariatem Boosting auf die einzelnen Komponenten der Zielgrösse, was darlegt, dass der multivariate Ansatz sehr nützlich sein kann.

In der dritten Erweiterung werden fünf Robustifizierungen von L_2-Boosting für lineare Regression mit verschiedenen Robustheitseigenschaften betrachtet. Die ersten beiden verwenden den Huber Verlust als implementierende Verlustfunktion für Boosting und die nächsten beiden verwenden robuste einfache lineare Regression für das anpassen beim L_2Boosting (d.h. robuste Basis Lerner). Beide Konzepte können mit oder ohne Heruntergewichten von Hebelpunkten angewendet werden. Die letzte Methode verwendet robuste Korrelationsschätzer und scheint die am meisten robuste zu sein. Entscheidende Vorteile aller Methoden sind, dass sie keine Kovarianzmatrizen aller Kovariablen berechnen und dass sie nicht multivariate Hebelpunkte identifizieren müssen. Falls es keine Ausreisser gibt, sind die robusten Methoden nur wenig schlechter als L_2Boosting. Bei kontaminierten Daten übertreffen die robusten Methoden L_2Boosting jedoch deutlich. Einige der Robustifizierungen können sehr effizient berechnet werden und sind deshalb gut geeignet für hoch-dimensionale Probleme.

Schliesslich habe ich LogitBoost mit einem auf Bäumen basierenden Lerner auf die fünf Datensätze der "Performance Prediction Challenge" des "World Congress on Computational Intelligence (WCCI) 2006" angewendet. Die Anzahl Boosting Iterationen und die Grösse der Bäume werden mit 10-facher Kreuzvalidierung bestimmt. Eine ein-

fache “Shrinkage”-Strategie wurde hinzugefügt, um den Algorithmus stabiler zu machen. Die Resultate sind vielversprechend, habe ich doch den Wettbewerb gewonnen.