Friedman, Jerome; **Hastie, Trevor**; **Tibshirani, Robert**

Additive logistic regression: a statistical view of boosting. (With discussion and a rejoinder by the authors). (English) Zbl 1106.62323

Ann. Stat. 28, No. 2, 337-407 (2000).

Summary: Boosting is one of the most important recent developments in classification methodology. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced. For many classification algorithms, this simple strategy results in dramatic improvements in performance. We show that this seemingly mysterious phenomenon can be understood in terms of well-known statistical principles, namely additive modeling and maximum likelihood. For the two-class problem, boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion. We develop more direct approximations and show that they exhibit nearly identical results to boosting. Direct multiclass generalizations based on multinomial likelihood are derived that exhibit performance comparable to other recently proposed multiclass generalizations of boosting in most situations, and far superior in some. We suggest a minor modification to boosting that can reduce computation, often by factors of 10 to 50. Finally, we apply these insights to produce an alternative formulation of boosting decision trees. This approach, based on best-first truncated tree induction, often leads to better performance, and can provide interpretable descriptions of the aggregate decision rule. It is also much faster computationally, making it more suitable to large-scale data mining applications.

MSC:

62G08 Nonparametric regression and quantile regression
62H30 Classification and discrimination; cluster analysis (statistical aspects)
68T05 Learning and adaptive systems in artificial intelligence

Keywords:

classification; tree; nonparametric estimation; stagewise fitting; machine learning

Software:

AdaBoost.MH

Full Text: DOI Euclid

References:

[1] Breiman, L. (1996). Bagging predictors. Machine Learning 24 123-140. · Zbl 0858.68080

[2] Breiman, L. (1997). Prediction games and arcing algorithms. Technical Report 504, Dept. Statistics, Univ. California, Berkeley. Breiman, L. (1998a). Arcing classifiers (with discussion). Ann. Statist. 26 801-849. Breiman, L. (1998b). Combining predictors. Technical report, Dept. Statistics, Univ. California, Berkeley. · Zbl 0934.62064

[3] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA. · Zbl 0541.62042

[4] Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). Ann. Statist. 17 453-555. · Zbl 0689.62029 · doi:10.1214/aos/1176347115

[5] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. Proc. IEEE Trans. Inform. Theory 13 21-27. · Zbl 0154.44505 · doi:10.1109/TIT.1967.1053964

[6] Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning? 1-22.

[7] Freund, Y. (1995). Boosting a weak learning algorithm by majority. Inform. and Comput. 121 256-285. Freund, Y. and Schapire, R. (1996a). Game theory, on-line prediction and boosting. In Proceedings of the Ninth Annual Conference on Computational Learning Theory 325-332. Freund, Y. and Schapire, R. E. (1996b). Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference 148-156. Morgan Kaufman, San Francisco. · Zbl 0833.68109

[8] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. J. Comput. System Sciences 55. · Zbl 0880.68103 · doi:10.1006/jcss.1997.1504

[9] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). Ann. Statist. 19 1-141. · Zbl 0765.62064 · doi:10.1214/aos/1176347963

[10] Friedman, J. (1996). Another approach to polychotomous classification. Technical report, Stanford Univ.

[11] Friedman, J. (1999). Greedy function approximation: the gradient boosting machine. Technical report, Stanford Univ.

[12] Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assoc. 76 817-823.

[13] Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall, London. · Zbl 0747.62061

[14] Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. Ann. Statist. 26 451-471. · Zbl 0932.62071 · doi:10.1214/aos/1028144844

[15] Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. J. Amer. Statist. Assoc. 89 1255-1270. · Zbl 0812.62067 · doi:10.2307/2290989

[16] Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning 11 63-90. · Zbl 0850.68278 · doi:10.1023/A:1022631118932

[17] Huber, P. (1964). Robust estimation of a location parameter. Ann. Math. Statist. 53 73-101. · Zbl 0136.39805 · doi:10.1214/aoms/1177703732

[18] Kearns, M. and Vazirani, U. (1994). An Introduction to Computational Learning Theory. MIT Press.

[19] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Processing 41 3397-3415. · Zbl 0842.94004 · doi:10.1109/78.258082

[20] McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. Chapman and Hall, London. · Zbl 0744.62098

[21] Schapire, R. (1997). Using output codes to boost multiclass learning problems. In Proceedings of the Fourteenth International Conference on Machine Learning 313-321. Morgan Kaufman, San Francisco.

[22] Schapire, R. E. (1990). The strength of weak learnability. Machine Learning 5 197-227.

[23] Schapire, R. E. and Singer, Y. (1998). Improved boosting algorithms using confidence-rated predictions. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory. · Zbl 0945.68194

[24] Schapire, R., Freund, Y., Bartlett, P. and Lee, W. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Statist. 26 1651-1686. · Zbl 0929.62069 · doi:10.1214/aos/1024691352

[25] Valiant, L. G. (1984). A theory of the learnable. Comm. ACM 27 1134-1142. · Zbl 0587.68077 · doi:10.1145/1968.1972

[26] in a paper by Amit and Geman (1997). Using this approach and 100 iterations gives the following test-set errors as compared to the best corresponding values for LogitBoost.

[27] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural Comput. 9 1545-1588. Breiman, L. (1999a). Random forests. Technical report, available at www.stat.berkeley.edu. Breiman, L. (1999b). Prediction games and arcing algorithms. Neural Comput. 11 1493-1517. URL:

[28] Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Mach. Learning 40 139-158.

[29] Wheway, V. (1999). Variance reduction trends on 'boosted' classifiers. Available from virg@cse. unsw.edu.au Friedman (1999). It does not involve any "reweighting". The weak learner b \cdot is fitted in the mth step to the current residuals Yi Fm-1 Xi There is no need for a surrogate loss function in this case since the evaluating L2 loss is the best possible for Newton's method and the quadratic approximation is phenomena are reported in Friedman (1999). From our own work [B ühlmann and Yu (2000)] we know that stumps evaluated at x have high variances for x in a whole region of the covariate space. From an asymptotic point of view, this region is "centered around" the true optimal split point for a stump and has "substantial" size O n-1/3 . That is, stumps do have high variances even in low dimensions as in this simple case (with only three parameters) as long as one is looking at the "right scale" O n-1/3 ; such a high variance presumably propagates when combining stumps in boosting. This observation is the starting point for another boosting machine to be described next. · Zbl 1213.62109

[30] Breiman, L. (1996). Bagging predictors. Machine Learning 24 123-140. · Zbl 0858.68080

[31] B ühlmann, P. and Yu, B. (2000). Explaining bagging.

[32] Friedman, J. (1999). Greedy function approximation: the gradient boosting machine. Technical report, Stanford Univ.

[33] Gill, E. P., Murray, W. and Wright, M. H. (1981). Practical Optimization. Academic Press, New York. · Zbl 0503.90062

[34] Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall, London. Seminar f ür Statistik ETH-Zentrum, LEO D72 CH-8092 Zurich Switzerland E-mail buhlmann@stat.math.ethz.ch Department of Statistics University of California Berkeley, California 94720-3860

[35] Breimann, L. (1996). Bagging predictors. Machine Learning 24 123-140. · Zbl 0858.68080

[36] Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assoc. 76 817-823.

[37] Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning 11 63-90. · Zbl 0850.68278 · doi:10.1023/A:1022631118932

[38] proposed by Freund and Mason (1999). They represent decision trees as sums of very simple functions and use boosting to simultaneously learn both the decision rules and the way to average them. Another important issue discussed in this paper is the performance of boosting methods on data which are generated by classes that have a significant overlap, in other words, classification problems in which even the Bayes optimal prediction rule has a significant error. It has been observed by several authors, including those of the current paper, that AdaBoost is not an optimal method in this case. The problem seems to be that AdaBoost overemphasizes the atypical examples which eventually results in inferior rules. In the current Freund (1999).

[39] Breiman, L. (1998). Arcing classifiers. Ann. Statist. 26 801-849. · Zbl 0934.62064 · doi:10.1214/aos/1024691079

[40] Freund, Y. (1999). An adaptive version of the boost by majority algorithm. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory. · Zbl 0988.68150

[41] Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. In Machine Learning: Proceedings of the Sixteenth International Conference 124-133.

[42] Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. Ann. Statist. 26 1651-1686. · Zbl 0929.62069 · doi:10.1214/aos/1024691352

[43] , takes on an interesting form when we have a sample.

[44] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36 105-139.

[45] Breiman, L. (1999). Using adaptive bagging to debias regressions. Technical Report 547, Dept. Statistics, Univ. California, Berkeley. · Zbl 1052.68109

[46] Chipman, H., George, E. and McCulloch, R. (1998). Bayesian CART model search (with discussion). J. Amer. Statist. Assoc. 93 935-960.

[47] Denison, D. (2000). Boosting with Bayesian stumps. Technical report, Dept. Mathematics, Imperial College.

[48] Denison, D., Mallick, B. and Smith, A. (1996). Bayesian CART. Technical report, Dept. Mathematics, Imperial College. · Zbl 1048.62502

[49] DiMatteo, I., Genovese, C. and Kass, R. (1999). Bayesian curve fitting with free-knot splines. Technical report, Carnegie Mellon Univ. · Zbl 0986.62026

[50] Drucker, H. and Cortes, C. (1996). Boosting decision trees. In Proceedings of Neural Information Processing 8 479-485. MIT Press.

[51] Elkan, C. (1997). Boosting and naïve Bayes learning. Technical Report CS97-557, Univ. California, San Diego.

[52] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55 119-139. · Zbl 0880.68103 · doi:10.1006/jcss.1997.1504

[53] Hastie, T. and Tibshirani, R. (1998). Bayesian backfitting. Technical report, Stanford Univ. · Zbl 1059.62524

[54] Heikkinen, J. (1998). Curve and surface estimation using dynamic step functions. In Practical Nonparametric and Semiparametric Bayesian Statistics (D. Dey, P. M üller and D. Sinha, eds.) 255-272. Springer, New York. · Zbl 0918.62031

[55] Lee, J. (1999). A computer program that plays a hunch. New York Times August 17.

[56] Quinlan, J. (1996). Bagging, boosting, and C4.5. In Proceedings Thirteenth American Association for Artificial Intelligence National Conference on Artificial Intelligence 725-730. AAAI Press, Menlo Park, CA. · Zbl 1184.68423

[57] Ridgeway, G. (1999). The state of boosting. In Proceedings of the Thirty-first Symposium on the Interface 172-181.

[58] Ridgeway, G. (1999). The state of boosting. In Computing Science and Statistics 31 (K. Berk, M. Pourahmadi, eds.) Interface Foundation of North America, 172-181. Fairfax, VA.

[59] Breiman, L. (1999). Using adaptive bagging to debias regressions. Technical Report 547, Dept. Statistics, Univ. California, Berkeley. · Zbl 1052.68109

[60] Freund, Y. (1999). An adaptive version of boost by majority algorithm. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory. · Zbl 0988.68150

[61] Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. In Machine Learning: Proceedings of the Sixteenth International Conference 124-133.

[62] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). Ann. Statist. 19 1-141. Friedman, J. H. (1999a). Greedy function approximation: a gradient boosting machine. Ann. Statist. To appear. Friedman, J. H. (1999b). Stochastic gradient boosting. Technical report, Dept. Statistics, Stanford Univ. · Zbl 0765.62064 · doi:10.1214/aos/1176347963

[63] Friedman, J. H. and Hall, P. (1999). On bagging and nonlinear estimation. J. Comput. Graph. Statist. · Zbl 1104.62047

[64] Grove, A. and Schuurmans, D. (1998). Boosting in the limit: maximizing the margin of learned ensembles. In Proceedings of the Fifteenth National Conference on Artificial Intelligence.

[65] Quinlan, J. (1996). Boosting first order learning. In Proceedings of the Seventh International Workshop on Algorithmic Learning Theory (S. Arikawa and A. Sharma, eds.) Lecture Notes in Artificial Intelligence 1160 143-155. Springer, Berlin. · Zbl 1184.68423

[66] Ratsch, G. (1998). Ensemble learning methods for classification. Masters thesis, Dept. Computer Science, Univ. Potsdam.

[67] Ratsch, G., Onoda, T. and Muller, K. R. (2000). Soft margins for AdaBoost. Machine Learning 1-35. · Zbl 0969.68128

[68] Ridgeway, G. (1999). The state of boosting. In Proceedings of the Thirty-first Symposium on the Interface 172-181.