A Fast Ensemble of Regressors

M. Karagiannopoulos, D. Anyfantis, S. B. Kotsiantis, P. E. Pintelas

Educational Software Development Laboratory
Department of Mathematics, University of Patras, Hellas

Abstract

We have implemented a learning tool that combines the RepTree, the linear regression and the Decision Stump algorithms using the averaging methodology. We performed a large-scale comparison with other state-of-the-art algorithms and fast ensembles on several datasets and we took better accuracy in most cases using less time for training, too.

Key words: machine learning, regression, data mining

1 Introduction

Combining regressors is proposed as a new direction for the improvement of the accuracy of regression models [3]. However, ensembles need increased computation and a research area is to explore learning techniques for scaling up to large datasets. In this work, we try to bridge the gap by using fast weak algorithms for building a rapid ensemble. Section 2 discusses the proposed ensemble method and experiment results of the proposed ensemble with other learning. We conclude in Section 3.

2 Proposed Ensemble

The training time is often less for generating multiple weak regressors compared to training one strong regressor. This is because strong regressors spend a majority of their training time in fine tuning. Secondly, weak regressors are

Email address: {mariosk, dany, sotos, pintelas}@math.upatras.gr (M. Karagiannopoulos, D. Anyfantis, S. B. Kotsiantis, P. E. Pintelas).

also less likely to suffer from overfitting problems. As far as the used learning algorithms of the proposed ensemble are concerned, three fast algorithms are used: 1) Linear regression (LR) [6], 2) RepTree [11] and 3) Decision stumps (DS) [8]. The corresponding predictions of the base regression models are then combined with averaging rule to produce the final decision. It must be also mentioned that the proposed ensemble can be easily distributed and parallelized. This parallel and distributed execution of the presented ensemble can achieve linear speedup. For our study, we used a number of well-known datasets by many domains from the UCI repository [1]. In order to calculate the models' correlation coefficient for our experiments, cross validation was run 10 times for each algorithm and the average value was calculated. It must be mentioned that we used the free available source code for most of the algorithms by [11] for our experiment. During the experiment, the proposed ensemble was compared with a representative algorithm for each of the other sophisticated machine learning techniques: Back Propagation (BP) algorithm [11], SMOreg algorithm [5], Kstar algorithm [9] and decision table algorithm [10]. In Table 1 and Table 2, we represent with "v" that the proposed ensemble looses from the specific algorithm. That is, the specific algorithm performed statistically better than the proposed according to paired t-test with p¡0.01. Furthermore, in Tables, "*" indicates that proposed ensemble performed statistically better than the specific regressor according to paired t-test with p¡0.01. In all the other cases, there is no significant statistical difference between the results (Draws). We also compare the proposed ensemble with other fast ensembles: Bagging RepTree, Bagging DS, Boosting RepTree and Boosting DS. Bagging is a method for building ensembles that uses different subsets of training data with a single learning method [3]. Additive Regression [7] is a practical implementation of the boosting [4].

3 Conclusion

The proposed ensemble needed less time for training than all the tested algorithms. The proposed ensemble can also achieve an increase in correlation coefficient from 2% to 17% compared to other learners. In a future work, the proposed ensemble will be made agent-based.

Table 1. Comparing the proposed ensemble with well known regressors

	AverageLRD	Kstar	DT	BP	SMOreg
auto93	0.80	0.77*	0.68*	0.85v	0.82
autoHorse	0.92	0.90	0.85*	0.95v	0.95v
autoMpg	0.90	0.91	0.90	0.91	0.92

autoPrice	0.90	0.91	0.81*	0.90	0.90
bodyfat	0.97	0.87*	0.97	0.98	0.99
breastTumor	0.27	0.19*	0.16*	0.09*	0.28
cholesterol	0.16	0.04*	0.07*	0.08*	0.16
cpu	0.97	0.97	0.92*	1.00v	0.97
echoMonths	0.71	0.39*	0.72	0.42*	0.68*
elusage	0.85	0.85	0.88v	0.86	0.84
hungarian	0.68	0.55*	0.59*	0.49*	0.58*
lowbwt	0.79	0.62*	0.78	0.60*	0.77
pbc	0.57	0.30*	0.40*	0.32*	0.58
pwLinear	0.89	0.72*	0.83*	0.90	0.86*
quake	0.10	0.08	0.09	0.08	0.06*
sensory	0.47	0.39*	0.57v	0.29*	0.35*

Table 2. Comparing the proposed ensemble with well known regressors

	AverageLRD	Bagging-RepTree	Boosting-RepTree	Bagging-DS	Boosting-DS
auto93	0.80	0.43*	0.26*	0.74*	0.79
autoHorse	0.92	0.89*	0.85*	0.80*	0.90
autoMpg	0.90	0.91	0.89	0.78*	0.90
autoPrice	0.90	0.92	0.90	0.82*	0.91
bodyfat	0.97	0.98	0.98	0.84*	0.97
breastTumor	0.27	0.22*	0.16*	0.23*	0.29
cholesterol	0.16	0.18	0.07*	0.12*	0.14
cpu	0.97	0.96	0.90*	0.87*	0.97
echoMonths	0.71	0.69	0.69	0.69	0.59*
elusage	0.85	0.82*	0.80*	0.84	0.83
hungarian	0.68	0.64*	0.58*	0.60*	0.67
lowbwt	0.79	0.79	0.77	0.78	0.77

pbc	0.57	0.55	0.46*	0.46*	0.53*
pwLinear	0.89	0.91	0.90	0.68*	0.85*
quake	0.10	0.12	0.06	0.09	0.08

References

[1] C. Blake & C. Merz, UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/ mlearn/MLRepository.html

[2] Breiman L., Bagging Predictors. Machine Learning 24 (1996) 123-140.

[3] Gavin Brown, Jeremy Wyatt, and Peter Tino. Managing diversity in regression ensembles. Journal of Machine Learning Research, 6, 2005.

[4] Chan Ph., Fan W., Prodromidis A., Stolfo, S., Distributed Data Mining in Credit Card Fraud Detection, IEEE Intelligent Systems on Data Mining, December 1999.

[5] Duffy, N. Helmbold, D., Boosting Methods for Regression, Machine Learning, 47, 153-200, 2002

[6] Gary William Flake, Steve Lawrence, Efficient SVM Regression Training with SMO, Machine Learning, Volume 46, Issue 1 - 3, Jan 2002, Pages 271 - 290.

[7] Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, ISBN: 080394540X, Sage Pubns.

[8] Friedman J. (2002). "Stochastic Gradient Boosting," Computational Statistics and Data Analysis 38(4):367-378.

[9] Iba, W., & Langley, P., Induction of one-level decision trees. Proceedings of the Ninth International Machine Learning Conference (1992). Aberdeen, Scotland: Morgan Kaufmann.

[10] C. John and L. Trigg, K*: An Instance- based Learner Using an Entropic Distance Measure", Proc. of the 12th International Conference on ML, pp. 108-114, 1995.

[11] Kohavi R. (1995). "The Power of Decision Tables." In Proc European Conference on Machine Learning.

[12] Mitchell, T., Machine Learning. McGraw Hill (1997).

[13] Shevade, S., Keerthi, S., Bhattacharyya C., and Murthy, K. (2000). Improvements to the SMO algorithm for SVM regression. IEEE Transaction on Neural Networks, 11(5):1188-1183.

[14] Witten I. & Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo (2000).