ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Using Regression Error Characteristic Curves for Model Selection in Ensembles of Neural Networks

Aloísio Carlos de Pina and Gerson Zaverucha *

Universidade Federal do Rio de Janeiro, COPPE/PESC
Department of Systems Engineering and Computer Science
C.P. 68511 - CEP. 21945-970, Rio de Janeiro, RJ - Brazil

Abstract. Regression Error Characteristic (REC) analysis is a technique for evaluation and comparison of regression models that facilitates the visualization of the performance of many regression functions simultaneously in a single graph. The objective of this work is to present a new approach for model selection in ensembles of Neural Networks, in which we propose the use of REC curves in order to select a good threshold value, so that only residuals greater than that value are considered as errors. The algorithm was empirically evaluated and its results were analyzed also by means of REC curves.

1 Introduction

An ensemble [1] is a set of predictors whose individual decisions are combined in some way. Ensembles are often much more accurate than the individual predictors that make them up. Many methods for constructing ensembles have been developed [1, 2]. Boosting [3] is one of the most popular methods for constructing ensembles. It produces multiple models by repeatedly altering the set of training examples given to a learner. Examples are given weights, and at each iteration a new hypothesis is learned and the examples are reweighted to focus the system on examples that the latest hypothesis gets wrong. The final predictor is constructed by a weighted vote of the individual models. Each predictor is weighted according to its accuracy. In the past few years, several studies were carried through about boosting [4].

In [5] it is presented a boosting algorithm for regression called Regressor-Boosting algorithm, based on a fundamental observation that often the mean squared error (MSE) of a predictor is significantly greater than the squared median of the error, due to a small number of big errors. By reducing the number of big errors, it is possible to reduce the MSE. The problem is to define a threshold from where a residual must be considered as a big error, since several factors are involved.

The objective of this work is to present a new approach for model selection in ensembles of Neural Networks, based on the Regressor-Boosting algorithm, in which we propose the use of the Regression Error Characteristic (REC) curves [6] in order to select a good threshold value, so that only residuals greater than that value are considered as errors. The REC curves facilitate visual comparison of regression functions and they are qualitatively invariant to choices of error metrics and scaling.

* The authors are partially supported by the Brazilian Research Agency CNPq.

Besides, the curve area provides a valid measure of the expected performance of the regression model and the information represented in REC curves can be used to guide the modeling process. An extensive experimental evaluation was carried through in order to analyze the performance of the new approach.

This paper is organized as follows. The next section has a brief review of the main characteristics of the Regressor-Boosting algorithm. In Section 3, a summary of REC curves is presented. Then, the boosting algorithm for regression using REC curves is explained in Section 4. In Section 5, an experimental evaluation of the new boosting approach is reported. Finally, in Section 6, the conclusions and the plans for future research are presented.

2 The Regressor-Boosting Algorithm

The essence of the algorithm is the construction of an ensemble of three estimators. They are trained on different input distributions and their outputs are combined in order to reduce the big error rate. An error is considered big if it is greater than a threshold value γ (it is then called big error with reference to γ).

The algorithm works as follow. The training data is split in three sets. The first set is used to train a first expert. The second set is used to build the training set for the second expert. This second training set contains the instances from the second set on which the first expert has a big error and a similar amount of instances on which it does not have a big error. The two experts are tested with the third set, and the training set of the third expert contains only those instances from the third set on which exactly one of the two first experts had a big error. The output of the ensemble is the median of the outputs of the three experts.

Instead of the majority vote of an ensemble used for classification tasks, the combination model used is the median of the three predictors. When the ensemble is used for prediction the worst of the three estimates for any sample point is irrelevant, because the median must be one of the other estimates. It was shown that the median is theoretically and empirically better than the average of the outputs [5] and may also be applied to more recent versions of the boosting algorithm.

3 Regression Error Characteristic Curves

Results achieved by Provost, Fawcett and Kohavi [7] and indicate ROC analysis [8] as a superior methodology than the accuracy comparison in the evaluation of classification learning algorithms. But ROC curves are limited to classification problems. Regression Error Characteristic (REC) curves [6] generalize ROC curves to regression with similar benefits.

The REC curve is a technique for evaluation and comparison of regression models that facilitates the visualization of the performance of many regression functions simultaneously in a single graph. An REC graph contains one or more monotonically increasing curves (REC curves) each corresponding to a regression model. One can easily compare many regression functions by examining the relative position of their REC curves. The shape of the curve reveals additional information that can be used to guide modeling.

REC curves plot the error tolerance on the x-axis and the accuracy of a regression function on the y-axis. Accuracy is defined as the percentage of points predicted within the tolerance. The area over the REC curve (AOC) is an estimate of the expected error for a regression model. The smaller the AOC is, better the regression function will be.

In order to adjust the REC curves in the REC graph, a null model is used to scale the REC graph. Reasonable regression approaches produce regression models that are better than the null model. The null model can be, for instance, the mean model: a constant function equal to the mean of the response of the training data. An example of REC graph can be seen in Fig. 1. The number between parentheses in the figure is the AOC value for each REC curve.

4 Boosting for Regression Using REC Curves

The Regressor-Boosting Algorithm uses a threshold γ for big errors. In practice there are several considerations when choosing a value for γ, such as the size of the data set. The optimal γ is one for which the big errors are responsible for a significant part of the MSE, but the big error rate is low. Usually the sets on which the second and third estimators are trained are more difficult and have a higher big error rate. In most cases, the choice of a good γ may require tuning.

We propose the use of REC curves in order to find a good value for γ. We select the value for the threshold γ as function of the AOC of the REC curve obtained for each expert. The big errors are then those residuals greater than f(AOC), where f(.) must be defined. We achieved good results by simply multiplying the AOC value by a scalar, adjusted by means of a validation set in an internal cross-validation [9].

Nevertheless, the best improvement provided by this approach is that we can access all the benefits of the REC curves described in Section 3, making possible a better evaluation of each part of the ensemble and facilitating the tuning.

5 Experimental Evaluation

In the experiments, 20 data sets were used in order to include several domains and difficulties. The data sets have been obtained from the UCI Machine Learning Repository [10], Delve repository (http://www.cs.toronto.edu/~delve/), Luís Torgo's Home Page (http://www.niaad.liacc.up.pt/~ltorgo/), Brazilian utilities [11], MLnet Archive (http://www.mlnet.org/) and StatLib data (http://lib.stat.cmu.edu/). A summary of the main characteristics of the used data sets can be found in [12]. The test method used in this research was the 10-fold cross-validation [13].

The implementation of the Neural Networks [14] used as experts (multi-layer perceptrons) was obtained from the WEKA System (http://www.cs.waikato.ac.nz/~ml/weka/). The network uses backpropagation to train [15]. The hidden layer is composed by n nodes with sigmoid activation function, where n is equal to half of the number of attributes. The output is a single unthresholded linear unit. The learning rate is equal to 0.3. In order to avoid overfitting, the number of epochs to train was adjusted by means of an internal cross-validation with validation sets.

We tested both methods for combining the outputs of the three experts: the median and the average. Some tests confirmed the results achieved by Avnimelech and Intrator [5], in which the median showed to be the best approach. Fig. 2 shows one of these cases. However, in the majority of tests the difference was negligible. We also verified that the performance of the boosting is better or as good as that of the best expert. Fig. 3 illustrates this conclusion.

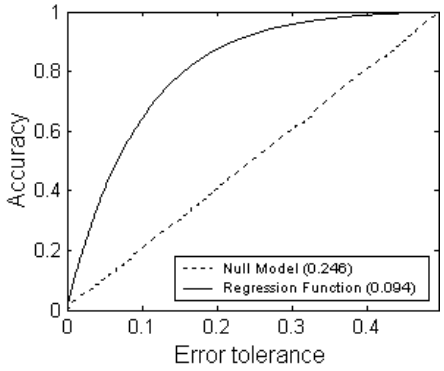

Fig. 1: Example of REC graph.

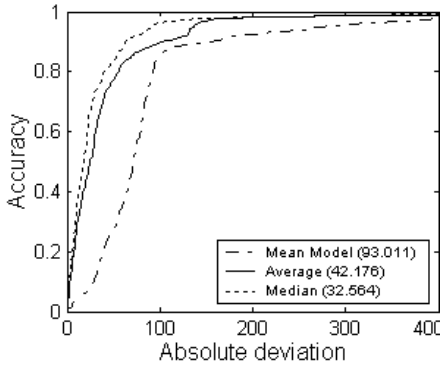

Fig. 2: Median vs. average as methods for combining the outputs of the three experts for the Machine-Cpu data set.

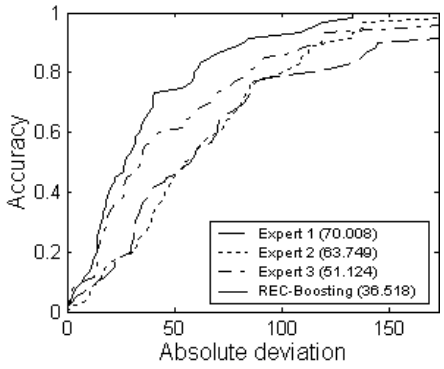

Fig. 3: Comparison of the individual performances of the experts and the boosting for the Pollution data set.

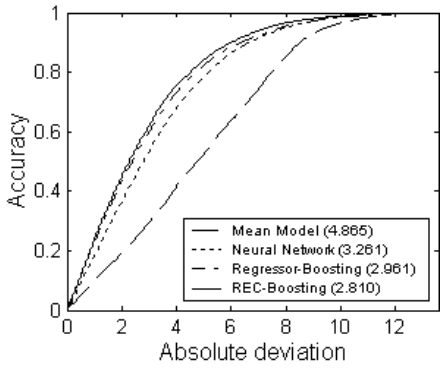

Fig. 4: Comparison of the models for the Pumadyn data set.

We have compared the Boosting for Regression using REC curves algorithm to the Regressor-Boosting algorithm, to a single Neural Network and to the mean model. The Neural Network was trained with the complete training set (the union of Set1, Set2 and Set3). The AOCs of the REC curves provided by each model for each data set are shown in Table 1. A summary of the results is presented in Table 2. The number of wins consists in to count the number of data sets where the REC-Boosting algorithm achieved a higher performance than a second model and to count the

number of data sets where it achieved a lower performance. The number of significant wins works in the same way, but a win is counted only if it is statistically significant according to the t-test at the 0.05 significance level. The last measure of performance considered is the score, where all models are compared simultaneously: for each data set, each model received a number of points between 0 and 3 (the best model received 3, the second place received 2 and so on until 0). The final value for each model is the ratio between the sum of all points received and the maximum number of points possible. Considering all the measures of performance, the Boosting for Regression using REC curves outperformed the other models.

For some data sets, the improvement in the results provided by the new boosting approach is clearly noted, as can be seen, for instance, in Fig. 4. Note that, although the REC curves are near, the curve corresponding to the Boosting for Regression using REC curves covers almost completely the others, thus the regression function generated by the ensemble dominates the other functions and therefore it is preferable.

Data Set	Mean Model	Neural Network	Regressor-Boosting	REC-Boosting
Abalone	2.111	1.577	1.729	1.527
Airplane Companies	5.440	5.559	5.185	6.627
Auto Imports	4488.607	2886.862	2176.815	2095.494
Auto-Mpg	6.475	4.331	3.819	4.273
Bank	0.124	0.027	0.025	0.023
Basketball Points	0.083	0.094	0.080	0.091
Brazilian utilities	0.056	0.062	0.054	0.058
Breast Cancer	29.146	34.471	30.907	31.496
California Housing	91144.647	56490.332	49086.957	53271.561
Census	32384.742	34602.705	24861.507	21257.289
Computer Activity	10.443	2.407	2.638	2.916
Elevators	0.0044	0.0018	0.0018	0.0017
Housing	6.566	6.837	5.986	5.946
Kinematics	0.216	0.124	0.106	0.109
Machine-Cpu	93.011	44.546	44.544	32.564
Pole Telecomm	32.211	10.465	10.547	9.328
Pollution	48.050	43.981	54.374	36.518
Pumadyn	4.865	3.261	2.961	2.810
PwLinear	3.566	1.402	1.896	1.986
Triazines	0.116	0.116	0.123	0.110

Table 1: AOCs of the REC curves provided for each data set.

Measure	Mean Model	Neural Network	Regressor-Boosting	REC-Boosting
Number of wins	16-4	17-3	11-9	-
Number of significant wins	14-0	7-1	5-2	-
Score	21.7	40.0	65.0	73.3

Table 2: Summary of results.

6 Conclusions and Future Works

We have presented here a boosting algorithm for regression that uses Regression Error Characteristic curves in order to define a good threshold for what we can consider as an error. Experimental tests have demonstrated the efficacy and applicability of the approach. By analyzing the REC curves and three measures of performance, we could verify that the Boosting for Regression using REC curves outperformed all models tested. Currently we are carrying through tests with model selection based on REC curves in ensembles of Neural Networks for time-series forecasting. As future works, we intend to investigate further the relation between the AOC and the performance of the ensemble and carry through comparisons against other ensemble regressors and other criteria to select the threshold.

References

[1] T. G. Dietterich, Machine Learning Research: Four Current Directions, *The AI Magazine*, 18:97-136, 1998.

[2] R. Caruana and A. Niculescu-Mizil, An Empirical Evaluation of Supervised Learning for ROC Area. In *Proceedings of the 1st Workshop on ROC Analysis in AI* (ROCAI 2004), pages 1-8, 2004.

[3] R. E. Schapire, The strength of weak learnability. *Machine Learning*, 5:197-227, 1990.

[4] R. E. Schapire, The boosting approach to machine learning: An overview. In *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[5] R. Avnimelech and N. Intrator, Boosting Regression Estimators, *Neural Computation*, 11:491-513, 1999.

[6] J. Bi and K. Bennett, Regression Error Characteristic Curves. In *Proceedings of the 20th International Conference on Machine Learning* (ICML 2003), pages 43-50, Washington, DC, 2003.

[7] F. Provost, T. Fawcett and R. Kohavi, The Case Against Accuracy Estimation for Comparing Classifiers. In *Proceedings of the 15th International Conference on Machine Learning* (ICML 1998), pages 445-453, Morgan Kaufmann, 1998.

[8] F. Provost and T. Fawcett, Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 43-48, AAAI Press, Newport Beach, CA, 1997.

[9] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[10] C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases, Machine-readable data repository, Department of Information & Computer Science, University of California, Irvine, 2005. [http://www.ics.uci.edu/~mlearn/MLRepository.html]

[11] M. Teixeira and G. Zaverucha, Fuzzy Bayes and Fuzzy Markov Predictors, *Journal of Intelligent and Fuzzy Systems*, 13:155-165, 2003.

[12] A. C. de Pina and G. Zaverucha, Boosting for Regression Using Regression Error Characteristic Curves. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning* (ROCML 2005), 2005.

[13] T. G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, *Neural Computation*, 10:1895-1924, 1998.

[14] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1998.

[15] D. E. Rumelhart and J. L. McClelland, *Parallel distributed processing: exploration in the microstructure of cognition, Vols. 1 & 2*, MIT Press, Cambridge, MA, 1986.