ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Ensemble Learning with Dynamic Ordered Pruning for Regression

Kaushala Dias and Terry Windeatt

Centre for Vision Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey, Guildford, Surrey, GU2 7XH
United Kingdom

Abstract. A novel method of introducing diversity into ensemble learning predictors for regression problems is presented. The proposed method prunes the ensemble while simultaneously training, as part of the same learning process. Here not all members of the ensemble are trained, but selectively trained, resulting in a diverse selection of ensemble members that have strengths in different parts of the training set. The result is that the prediction accuracy and generalization ability of the trained ensemble is enhanced. Pruning heuristics attempt to combine accurate yet complementary members; therefore this method enhances the performance by dynamically modifying the pruned aggregation through distributing the ensemble member selection over the entire dataset. A comparison is drawn with Negative Correlation Learning and a static ensemble pruning approach used in regression to highlight the performance improvement yielded by the dynamic method. Experimental comparison is made using Multiple Layer Perceptron predictors on benchmark datasets.

1 Introduction

It is recognized in the context of ensemble methods, the combined outputs of several predictors generally give improved accuracy compared to a single predictor [1]. Further performance improvements have also been shown by selecting ensemble members that are complementary [1]. The selection of ensemble members, also known as pruning, has the potential advantage of both reduced ensemble size as well as improved accuracy. However the selection of classifiers, rather than regressors, has previously received more attention and given rise to many different approaches to pruning [3]. Some of these methods have been adapted to the regression problem [3]. The proposed novel dynamic method, Ensemble Learning with Dynamic Ordered Pruning (ELDOP) for regression, uses the Reduced Error pruning method without back fitting (Section 3) for selecting the diverse members in the ensemble and only these are used for training [5]. To enhance the diversity, the selection and training of ensemble members are performed for every pattern in the training set.

By dynamic, we mean that the subset of predictors is chosen differently depending on its performance on the test sample. Given that only selected members of the ensemble are allowed to train for a given training pattern, the assumption is made that only a subset of the ensemble will perform well on a test sample. Therefore the method aims to automatically harness the ensemble diversity as a part of ensemble training. ELDOP is novel, since pruning occurs with training, and unlike [9], in the test phase there is no need to search for the closest training pattern.

2 Related Research

The main objective of using ensemble methods in regression problems is to harness the complementarity and diversity of individual ensemble member predictions [1]. In [2] ordered aggregation pruning using Walsh coefficient has been suggested. In Negative Correlation Learning, diversity of the predictors is introduced by simultaneously training a collection of predictors using a cost function that includes a correlation penalty term [6]; thereby collectively enhancing the performance of the entire ensemble. Empirical evidence shows that this approach tends to over-fit, but with an additional regularization term, Multi-objective Regularized Negative Correlation Learning tackles over-fitting for noisy data. By weighting the outputs of the ensemble members before aggregating, an optimal set of weights is obtained in [8] by minimizing a function that estimates the generalization error of the ensemble; this optimization being achieved using genetic algorithms. With this approach, predictors with weights below a certain level are removed from the ensemble. A dynamic ensemble selection approach in which many ensembles that perform well on an optimization set or a validation set are searched from a pool of over-produced ensembles and from this the best ensemble is selected using a selection function for computing the final output for the test sample [7]. In [9], for ordered aggregation, dynamically selecting the ensemble order that has been defined by the ensemble member performance on the training set has shown to improve prediction accuracy; here the ensemble order of the training pattern closest to the test pattern is searched and selected for the prediction phase. Here scaling factors come into effect when searching large training sets. Through instance selection [4], the training set is reduced by removing redundant or non-useful instances which improve prediction accuracy. The techniques used in instance selection can also be useful in pruning to design ensembles with improved diversity [5].

3 Reduced Error Pruning

Reduced Error Pruning without back fitting method (RE) [3], modified for regression problems, is used to establish the order of predictors in the ensemble that produces a minimum in the ensemble training error. Starting with the predictor that produces the lowest training error, the remaining predictors are subsequently incorporated one at a time into the ensemble to achieve a minimum ensemble error. The sub ensemble S_u is constructed by incorporating to S_{u-1} the predictor that minimizes

$$s_u = \arg_k \min u^{-1} \left(\sum_{i=1}^{u-1} C_{s_i} + C_k \right) \tag{1}$$

where for M number of predictors, $k \in (1,...,M) \backslash \{S_1, S_2, \ldots, S_{u-1}\}$ and $\{S_1, S_2, \ldots, S_{u-1}\}$ label predictors that have been incorporated in the pruned ensemble at iteration u-1. For the proposed method C_i is calculated per individual training pattern and expressed as

$$C_i = f_i(x_n) - y_n \tag{2}$$

where i = 1,2,…,M . The function $f_i(x)$ is the output of the i^{th} predictor and (x_n, y_n) is the training data where n = (1,2,…,N) training patterns. Therefore the information required for the ordering of the training error is contained in the vector C.

4 Method

Dynamic selection of ensemble members provides an ensemble tailored to the specific test instance. The method described here is for a regression problem where the ensemble members are simultaneously ordered and trained on a pattern by pattern basis. The ordering of ensemble members is based on the method of RE and only the first 50% of the ordered members for a given training pattern are used for learning. Therefore diversity is encouraged by training half of the ensemble members that perform well. The training continues until a pre-determined number of epochs of the training set are completed.

```
Training data $D = (x_n, y_n)$, where $n$ = (1,2,..,N) and $f_m$ is an ensemble member,
where $m$ = (1,2,..,M). $S$ is a vector with max index of $m$.
For $n$ = 1….N
    $S$ ← empty vector
    For $m$ = 1…M
        Evaluate $C_m = f_m(x_n) - y_n$
    End for
    For $u$ = 1…M
        min ← +∞
        For $k$ in (1,...,M)\{$S_1, S_2, …, S_u$}
            Evaluate $z = u^{-1}\left(\sum_{i=1}^{u-1} C_{S_i} + C_k\right)$
            If $z$ < min
                $S_u$ ← $k$
                min ← $z$
            End if
        End for
    End for
    Apply update rule to first 50% of members in $S$
End for
```

Fig 1: Pseudo-code implementing the training process with ordered ensemble pruning per training pattern.

The implementation of the proposed dynamic method consists of two stages. First the base ensemble members M are ordered and trained on a pattern by pattern basis. As shown in the pseudo-code in figure 1, this is achieved by building a series of nested ensembles in which the ensemble of size u contains the ensemble of size u-1. Taking a single pattern of the training set, the method starts with an empty ensemble *S*, in step 2, and builds the ensemble order, in steps 6 to 15, by evaluating the training error of each predictor in M. The predictor that increases the ensemble training error least is iteratively added to *S*. This is achieved by minimizing *z* in step 9. Then the update rule is applied to the first 50% of the ordered ensemble member in *S*. Therefore in one epoch of training, the Back Propagation update rule would be applied a different number of times for each predictor, the more effective predictors being trained the most.

In the second stage, the ensemble output for each test pattern is evaluated. The assumption is made that the outputs of ensemble members that perform well for a test pattern would cluster together. Therefore the second stage starts by clustering the ensemble outputs into two clusters. This is shown in step 1 in figure 2. Then in step 2, the mean and the standard deviations are calculated. Taking the ensemble member outputs of each cluster, the outputs that are within one standard deviation from the mean are selected for the sub-cluster of each original cluster. This is denoted by S_k. Finally the mean of each of these sub-clusters are calculated as the outputs of the original clusters. This is shown in step 10. In this paper the cluster output that is close to the test pattern output is selected.

```
Ensemble member output for a test pattern ($x_n$, $y_n$) is $f_m$, where $m$ = (1,2,..,M).
$f_j$ are ensemble member outputs in cluster $C_k$, $j$ = 1,2,..,J number of members.
$\mu_k$, $\sigma_k$ are the mean and the standard deviation of $C_k$
$S_k$ is the sub-cluster in $C_k$

 1. Using K-means (K = 2) separate $f_m$ into two clusters $C_1$ , $C_2$
 2. Find mean and standard deviation of the two clusters; $\mu_1$, $\sigma_1$, $\mu_2$, $\sigma_2$,
 3. Calculate cluster mean as follows for each of the two clusters $C_1$ , $C_2$:
 4. For $k$ = 1,2
 5.    For $j$ = 1….J
 6.          If $(\mu_k + \sigma_k) > f_j > (\mu_k - \sigma_k)$
 7.          Then    $S_k \leftarrow f_j$
 8.          End if
 9.    End for
10.    Evaluate the mean of $S_k$ ; $\bar{\bar{\mu}}_k$(This is the cluster output for comparison)
11. End for
```

Fig 2: Pseudo-code implementing the ensemble output evaluation for test pattern.

5 Results

MLP architecture with 5 nodes in the hidden layer, as described in [3] has been selected in this experiment. The training/test data split is 70/30 percent, and 32 base predictors are trained with identical training samples. The Mean Squared Error (MSE) is used as the performance indicator for both training and test sets, and averaged over 10 iterations. Training is stopped after fifty epochs.

Table 1 shows MSE performance comparison of Negative Correlation Learning (NCL) [6], Ordered Aggregation (OA) [3], Dynamic Ensemble Selection and Instantaneous Pruning (DESIP) [9] and the proposed method of Ensemble Learning with Dynamic Ordered Pruning (ELDOP). In table 1, grayed results indicate the minimum MSE over the four methods for every dataset. It is observed that the majority of the lowest MSE values have been achieved by ELDOP. Figure 3 shows the comparison of the training and test error plots with ensemble size for NCL, DESIP and ELDOP. It is observed that pruned ensembles with ELDOP are more accurate with fewer members than the other methods.

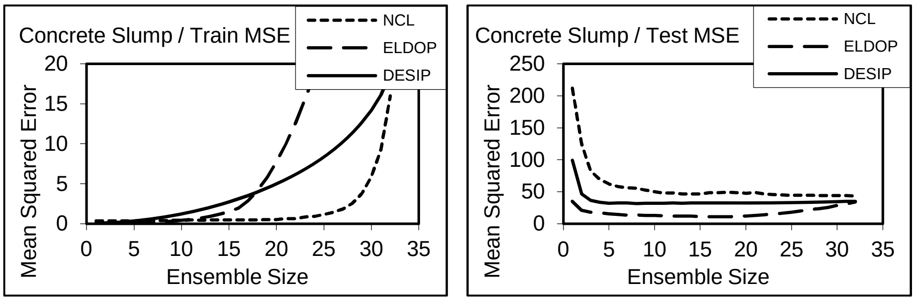

Fig 3: Comparison of the MSE plots of the training set and the test set for NCL, DESIP and ELDOP.

Dataset	Multiplier	NCL	OA	DESIP	ELDOP
Servo	10^0	0.25±0.49	1.35±1.69	0.14±0.24	0.10±0.14
Wisconsin	10^1	2.89±7.63	2.82±6.81	2.37±5.21	0.64±1.71
Concrete Slump	10^1	4.39±6.69	4.81±7.37	4.03±5.99	1.15±1.62
Auto93	10^2	0.52±1.57	1.02±2.73	0.72±1.92	0.45±1.50
Body Fat	10^1	0.10±0.34	3.66±4.62	0.09±0.32	0.29±0.52
Bolts	10^2	0.94±1.71	2.71±2.27	0.79±1.22	0.66±0.76
Pollution	10^3	1.99±3.38	3.57±5.56	1.70±2.68	2.14±3.19

Table 1: Averaged MSE with Standard Deviation for 10 iterations for NCL, OA, DESIP and ELDOP.

Dataset	Instances	Attributes	Source
Servo	167	5	UCI-Repository
Wisconsin	198	36	UCI-Repository
Concrete Slump	103	8	UCI-Repository
Auto93	82	20	WEKA
Body Fat	252	15	WEKA
Bolts	40	8	WEKA
Pollution	60	16	WEKA

Table 2: Benchmark datasets used

6 Conclusion

Unlike static ensemble pruning, dynamic pruning utilizes a distributed approach to ensemble selection and is an active area of research for both classification and regression problems. In this paper a novel method is introduced which combines ensemble learning with dynamic pruning of regression ensembles. Experimental results show that test error has been reduced by introducing pruning in the training phase of ensembles. In DESIP [9] the ensemble selection for a test pattern is based on the closest training instance and therefore a search is necessary to determine the pruned ensemble, while in ELDOP the ensemble is trained with the pruned selection, therefore eliminating the need to search. In NCL and DESIP the entire ensemble is utilized in training, while ELDOP trains only the selected members of the ensemble, with a commensurate reduction in training time. On a few datasets the proposed method has not improved performance, and will be investigated further along with methods that modify the cost function in NCL. Bias/Variance and time complexity analysis should also help to understand the performance relative to other ensemble methods with similar complexity.

References

[1] Tsoumakas G., Partalas I., Vlahavas I., An Ensemble Pruning Primer. Supervised and Unsupervised Ensemble Methods and their Applications. Studies in Computational Intelligence Volume 245, Springer 2009, pp 1 – 13.

[2] Windeatt T., Zor C., Ensemble Pruning using Spectral Coefficients. IEEE Trans. Neural Network. Learning Syst. 24(4), 2013. pp 673 – 678.

[3] Hernández-Lobato D., Martínez-Muñoz G., Suárez A, Empirical Analysis and Evaluation of Approximate Techniques for Pruning Regression Bagging Ensembles. Neurocomputing 74, 2011, pp 2250 – 2264.

[4] Olvera-Lopez J., Carrasco-Ochao J., Martinez-Trinidad J., Kittler J., A review of instance selection methods. Artificial Intelligence Review 34(2), Springer 2010, pp 133 – 143.

[5] Brown G., Wyatt J., Harris R., Yao X., Diversity creation methods: a survey and categorization. Information Fusion 6(1), 2005, pp 5 – 20.

[6] Chen H., Yao X., Multiobjective Neural Network Ensembles Based on Regularized Negative Correlation Learning. IEEE Trans. Knowledge and Data Engineering 22(12), 2010, pp1738 – 1751.

[7] Dos Santos E.M., Sabourin R., Maupin P., A Dynamic Overproduce-and-choose Strategy for the selection of Classifier Ensembles. Pattern Recognition 41, 2008, pp 2993 – 3009.

[8] Zhau Z.-H., Wu J., Tang W., Ensembling Neural Networks: many could be better than all, Artificial Intelligence, Volume 137, 2002, pp 239 – 263.

[9] Dias K., Windeatt T., Dynamic Ensemble Selection and Instantaneous Pruning for Regression. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2014, pp 643 – 648.