ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

Ensemble of hybrid networks with strong regularization

Shimon Cohen and Nathan Intrator*
School of Computer Science, Tel Aviv University

Ramat Aviv 69978, ISRAEL

Abstract. We introduce an ensemble of hybrid neural networks. The hybrid networks are composed of radial and projection units. They are trained using a deterministic algorithm that completely defines the parameters of the network for a given data set. Thus, there is no random selection of the initial (and final) parameters as in other training algorithms. Network independent is achieved by using an input sub-space random sampling as well as random selection of patterns from the training set. Few methods for ensemble fusion are explored and evaluated on several classification benchmark data-sets.

1 Introduction

Hybrid neural networks that are composed of radial basis functions and perceptrons have been recently introduced [5, 4]. Such networks employ a deterministic algorithm that computes the initial parameters from the training data. Thus, two networks that have been trained on the same data-set produce the same solution and therefore, a combination of such classifiers can not enhance the performance over a single one.

Fusion of experts has been studied extensively in recent years; One of the main results is that experts have to be partially independent for the fusion to be effective. The bagging algorithm can be used to de-correlate between classifiers as well as to obtain some performance measure on the accuracy of the classifiers using the "out of bag" sub-set of the data [1]. Another technique Arcing – adaptive re-weighting and combining – refers to reusing or selecting data in order to improve classification [2]. One popular arcing procedure is AdaBoost [9], in which the errors on the training data-sets are used to train more specific classifiers. Subsampling of the input space and the training patterns is extensively used in the random forest algorithm [3]. A different flavor of combination of classifiers use dynamic class combination (DCS) [10] and Classifiers Local Accuracy (CLA) in order to select the best classifier when making

*www.math.tau.ac.il/∼nin

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

a predication. This is done at the cost of saving the whole training set and then selecting the predication of the best classifier at the vicinity of a given pattern.

The hybrid Perceptron Radial Basis Function Network (PRBFN) is constructed in a very conservative manner and with strong regularization. It thus has a relatively small architecture and therefore a low variance [5, 4]. In this paper, we investigate the use of these methods on a combination of strong classifiers that has a deterministic training algorithm. Several ways to make the classifiers independent are considered as well as several combination strategies.

2 Training an ensemble

Since a PRBFN classifier is determined by the given training data set uniquely [4], independence in an ensemble can not come from random parameter initialization as in other algorithms [12]. This also implies that combination of such experts is obvisouly not optimal by a simple uniform averaging. The random forest algorithm uses sub-space resampling for each node in the tree [3]. AdaBoost uses a fraction of the data to train a classifier, thus different classifiers see different data-sets [9]. We use both techniques to make the classifiers more independent. We use subsample of the input space features, "boost" the classifiers and then combine the classifiers into an ensemble. We rely on the accuracy of the classifiers on the training set for experts fusion as explained below.

2.1 Ensemble generation

Given a data set $D = \{x_i, y_i\}_{i=1}^{N}$ where $x_i \in R^d$ and y_i is the class label. The input to the algorithm includes the maximum number of classifiers k_{max}, the size of the resampled subset of D is $n < N$ and $\gamma \in [0, 1]$ is the fraction of features for the random subspace selection.

- Initilialize: empty ensemble, $k = 0$
- while $k \leq k_{max}$
  - test the ensemble on the full training-set.
  - add to the current dataset D_k the misclassified patterns
  - select randomly $N - |D_k|$ from $D - D_k$ and add them to D_k
  - resample D_k on the features by using round$(\gamma * d)$ features.
- end-loop

The above algorithm differs from the AdaBoost algorithm [9], as **all** the misclassified patterns are added to the next subset (with probability 1). Since the PRBFN classifier is not a weak classifier, there are few such patterns, and thus, we include all the misclassified patterns in the data-set. Each classifier

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

receives a different part of the training data and a different subsample of the input variables as in Random Forests. Thus, dependency between experts is greatly reduced.

2.2 Experts fusion methodology

We have used three classifier combination rules. The first is the familiar majority vote; Here, the final decision is made by selection of the class with maximum number of votes in the ensemble.

The second strategy relies on a convex combination using the error values from the first stage of training. The output of the ensemble is given by:

$$f(x) = \sum_{k=1}^{M} a_k f_k(x) \qquad a_k \geq 0, \sum_{k=1}^{M} a_k = 1. \tag{1}$$

Let e_i be the classification error of the $i'th$ classifier. We set the weight of this classifier as follows:

$$a_k = \frac{\exp(-e_k)}{\sum_{i=1}^{M} \exp(-e_i)}, \tag{2}$$

where M is the number of classifiers in the ensemble. Motivated by the Gibbs distribution, which maximizes the information contribution from each classifier, the above equation simply gives stronger weight to classifiers with smaller error.

The third strategy involves dynamic selection of the best classifier for prediction of the output value when a novel pattern is given. When the confidence of the best classifier (to be explained below) is too low (below a given threshold) we use a dynamic combination of the classifiers to produce the output of the ensemble. We define a local accuracy for each classifier as follows. Let $k > 0$ and $x \in R^d$ be a novel pattern. Let $D_k(x)$ be the $k - nearest$ patterns in the training set to x. Set the local accuracy of the current classifier on x the to be:

$$l(x) = \frac{\sum_{x_j \in D_k(x)} \delta(arg\max_i p(y_i|x_j) - arg\max_j(t_j))}{k}, \tag{3}$$

where $\delta(x)$ is one for $x = 0$ and zero otherwise, and t_j is the target for pattern x_j. Thus, the local accuracy is the number of correct classified patterns in the $k - neighborhood$ of x. Let $l_1(x)$ be the maximum local accuracy and let $l_2(x)$ be the next highest accuracy. Define the confidence level as follows:

$$cl(x) = \frac{l_1(x) - l_2(x)}{l_1(x)}. \tag{4}$$

We further define the weights for each classifier as follows:

$$a_k(x) = \frac{\exp(l_k(x))}{\sum_{i=1}^{M} \exp(l_i(x))}, \tag{5}$$

Now the combination rule in this case is given by:

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

- Compute the local accuracy for each classifier as in 3.
- Compute the confidence level $cl(x)$ from Eq. 4.
- If $\max cl(x) > threshold$, select the output of the best classifier, otherwise use Eq. 1 where a_k is given by (5).

3 Results

We have considered several data-sets to evaluate the performance of the ensemble vs. PRBFN. The following methods of combination were used:

- ENS1-PRBFN The ensemble using the first selection (majority vote) strategy.
- ENS2-PRBFN The ensemble using a convex combination of classifiers where the errors affect the weight of the different classifiers in the ensemble.
- ENS3-PRBFN The ensemble using k-neighbors to select the best classifier.
- PRBFN the single classifier as described in [4].

Data sets description

The Breast-cancer dataset from the UCI repository was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. This dataset has 9 attributes and two classes and the number of training patterns is 699. The task is to classify the patterns to Benign or Malignant.

The Glass dataset from the UCI repository has 10 attributes and 7 types of glasses. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence if it is correctly identified! Ripley's best result on this data-set is 80% accuracy [11].

Method	Breast-cancer	Glass	Iris	Vowel	Pima
ENS1-PRBFN	96.5±1.4	96.2±3.5	95.3±4.5	85.2±5.3	77.4±3.2
ENS2-PRBFN	96.7±1.4	94.8±4.7	96.0±5.3	86.7±3.9	77.5±3.2
ENS3-PRBFN	96.8±1.9	94.2±5.2	96.0±4.6	87.3±4.0	77.4±4.4
PRBFN	96.0±2.03	93.3±5.7	95.3±4.6	81.8±5.9	76.6 ±6.0

Table 1: Comparison of percent correct classification rate of ensemble methods on several data-sets using 10 folds cross validation.

The Iris data-set [7] contains three classes, each with 50 instances. The classes refer to a type of iris plant. Each pattern is composed of four attributes.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

We used ten folds of cross validation in order to estimate the accuracy of the different classifiers.

The Deterding vowel recognition data [6, 8] is a widely studied benchmark. This problem may be more indicative of a real-world modeling problem. The data consists of auditory features of steady state vowels spoken by British English speakers. There are 528 training patterns and 462 test patterns. Each pattern consists of 10 features and belongs to one of 11 classes that correspond to the spoken vowel. The speakers are of both genders. This data, unlike the other data-sets that have been studied, has a fixed training and test set. Thus, we provide results with cross validation in Table 1, where we compare experts on cross validated test set and, for completeness, we provide an additional table with results on the fixed test set as is described in [6, 8]. Previous best score on the fixed test set was reported by Flake using SMLP units. His average best score was 60.6% [8] and was achieved with 44 hidden units. As can be seen in Table 2, the single PRBFN network surpasses this result and achieves 68.4% correct classification with only 22 hidden units. This result was achieved with a low variance architecture on a small training set, thus ensemble methods did not improve the result. When it is possible to decorrelate experts using a larger training set (Table 1), the ensemble improvement is more significant.

	ENS1-PRBFN	ENS2-PRBFN	ENS3-PRBFN	PRBFN
Vowel	66.1±1.3	67.8±1.4	67.5±0.9	68.1±0.0

Table 2: Classification rate on the Vowel data [6, 8] using the fixed test set with no cross validation.

4 Discussion

The performance of ensemble methods on a tight architecture which has been shown to have a low variance portion of the error was evaluated on several benchmark data-sets. Partial independence of the experts was achieved via boosting, and expert fusion was performed via majority (plurality), convex combination (as described above), and via dynamic fusion based on the local accuracy of each expert in the region close to the test pattern. We note that the improvement of ensemble of such architectures is smaller than improvement that can be achieved on other architectures which posses higher variance, nevertheless, improvement still exists, and is sometimes quite significant. The fusion methods we have studied do not appear to be significantly different in their improvement over a single expert. The key factor affecting the improvement is the degree of decorrelation of experts, which in this case, due to the deterministic nature of the architecture, depends on data resampling methods.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 337-342

References

[1] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[2] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

[3] L. Breiman. Random forests. Technical Report, Statistic Department University of California, Berkeley, 2001.

[4] S. Cohen and N. Intrator. Automatic model selection in a hybrid perceptron/radial network. *Information Fusion*, 3(4):259–266, 2002.

[5] S. Cohen and N. Intrator. A hybrid projection based and radial basis function architecture: Initial values and global optimization. *Pattern Anal. Appl. (Special issue on Fusion of Multiple Classifiers)*, 5(2):113–120, 2002.

[6] D.H. Deterding. *Speaker Normalisation for Automatic Speech Recognition.* PhD thesis, University of Cambridge, 1989.

[7] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[8] G.W. Flake. Square unit augmented, radially extended, multilayer percpetrons. In G. B. Orr and K. Müller, editors, *Neural Networks: Tricks of the Trade*, pages 145–163. Springer, 1998.

[9] Y. Freund and R.E. Schapire. A decision theorethic generalization of online learning and application to boosing. *Journal of Computer and System Sciences*, 55(1):119–139, 1995.

[10] G. Giacinto and F. Roli. Dynamic classifier selection. In *First International workshop on Multiple Classifier Systems*, pages 177–189, 2000.

[11] B. D. Ripley. *Pattern Recognition and Neural Networks.* Oxford Press, 1996.

[12] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group. Parallel distributed processing: Explorations in the microstructure of cognition. *The MIT Press*, 1986. Vol. 2.