ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Multi-View Forests of Tree-Structured Radial Basis Function Networks Based on Dempster-Shafer Evidence Theory

Mohamed Farouk Abdel Hady, Günther Palm, Friedhelm Schwenker*

University of Ulm, Department of Neural Information Processing
James Frank Ring, 89069 Ulm, Germany

Abstract. An essential requirement to create an accurate classifier ensemble is the diversity among the individual base classifiers. In this paper, *Multi-View Forests*, a method to construct ensembles of tree-structured radial basis function (RBF) networks using multi-view learning is proposed. In Multi-view learning it is assumed that the patterns to be classified are described by multiple feature sets (views). *Multi-view Forests* have been evaluated by using a benchmark data set of handwritten digits recognition. Results show that multi-view learning can improve the performance of the ensemble by enforcing the diversity among the individual classifiers.

1 Introduction

Error diversity is a fundamental requirement to build an effective classifier ensemble and therefore many definitions of classifiers diversity have been introduced e.g. ten different measures have been proposed by Kuncheva [1]. Multi-view learning is a machine learning approach where each pattern is represented by many feature sets obtained through different physical sources and sensors or derived by different feature extraction procedures leading to different types of discriminating information about the pattern. For example, a web page can be represented by different views, e.g. a distribution of words used in the web page, hyperlinks that point to this page, and any other statistical information, such as size, number of accesses, etc. The paper is organized as follows: In Section 2 the *Multi-view Forests* method, a new multi-view ensemble method, is explained. Results of its application to handwritten digits recognition are presented in Section 3 and finally we conclude the paper in Section 4.

2 Multi-View Forests

In the multi-view learning, the input space is given by a product space $X = X_1 \text{x} \ldots \text{x}\ X_F$ and data points are given by $x = (x_1, \ldots, x_F)$ where x_i denotes the i^{th} feature vector. A Multi-View Forest is an ensemble of tree-structured classifiers $\{TC_k\}$, $k=1,\ldots,M$ that are trained on the predefined subspaces $\{S_k\}$ (see Figure 1). The input patterns x are projected onto $\{S_k\}$. The tree classifier outputs $\{y_k\}$ are then combined to produce the final decision y_{Final} using a combination function such as minimum and product.

* This work has been supported by the German Science Foundation DFG under grant SCHW623/4-3 and a scholarship of the German Academic Exchange Service DAAD.

Tree-structured RBF networks are chosen as the base classifiers in the ensemble because of their ability to decompose multi-class recognition problems into less complex binary classification subtasks where each network is assigned a subtask.

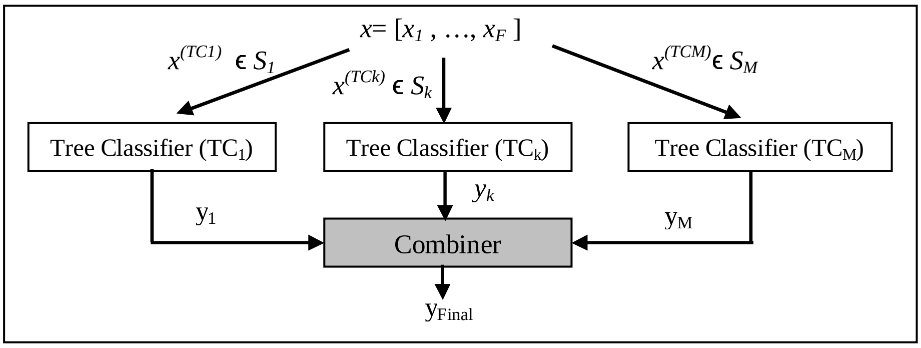

Fig. 1: Architecture of the proposed Multi-View Forest.

2.1 **Tree-Structured RBF Networks**

A tree classifier is an ensemble of K-1 embedded binary RBF networks (see Figure 3) solving a given K-class problem using a single feature set (Single-View Tree) or a group of feature sets (Multi-View Tree). In [2], Support Vector Machines have been used as binary classifiers to construct binary tree-structured classifier to solve multi-class problems.

2.1.1 Tree Training Phase

In the first step, the tree structure of the classifier is generated as follows: For each classifier node and for each feature space, k-means clustering of the class centroids was performed. At each node, the set of classes is splitted into two disjoint subsets until subset contains exactly one class (see Figure 2). The different splits are evaluated using a clustering evaluation measure. A scale and feature independent criterion to measure the compactness and separation quality of a pair of clusters is defined by the ratio of the inter-cluster distance to the sum of inner-cluster distances [3]. The best view is the one with the maximum ratio.

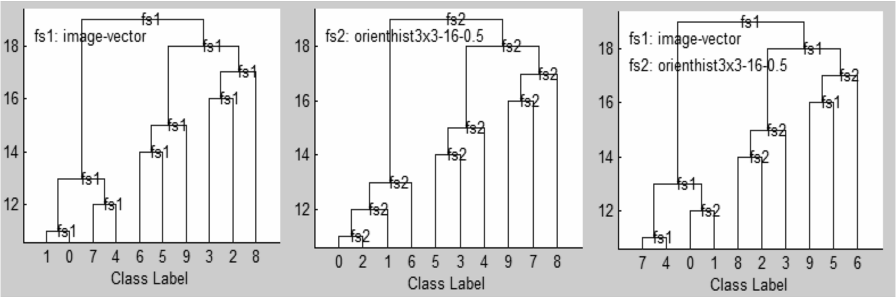

Fig. 2: Tree constructed for the digits data using the image vector and the Orientation Histogram feature vector.

In the second step of the training phase, an RBF network is assigned to each node (see Figure 3) where the Gaussian function in (1) is used as the radial basis function.

$$h_j(\|x-c_j\|)=\exp(-\|x-c_j\|^2/2\sigma_j^{\,2}) \qquad (1)$$

For training such an RBF network a two-phase learning procedure is used. In the first phase, the RBF centers $c_1, \ldots, c_k$ are determined by performing class-specific k-means clustering and the distance between c_j and the nearest prototype with a different class label is used as the width of the jth RBF neuron, [4].

$$\sigma_j=\alpha\min\{\ \|c_j-c_i\|:class(c_j)\neq class(c_i),i=1,\ldots,k\ \} \qquad (2)$$

Then, in the second phase the output layer weights $\boldsymbol{W}$ are computed directly by

$$HW=Y \qquad (3)$$

where $\boldsymbol{Y}$ is the matrix of target outputs of the $\boldsymbol{M}$ training examples and $\boldsymbol{H}$ is the activation matrix defined by

$$H=(h_j(\|x_i-c_j\|))_{j=1,\ldots,k}^{i=1,\ldots,M} \qquad (4)$$

Therefore, calculating the pseudoinverse of $\boldsymbol{H}$ provides a least squares solution to the system of linear equations in (3). This direct computation is a fast method yielding good classification results.

2.1.2 Tree Classification Phase

Two different strategies to compute the decision of the tree classifier have been evaluated throughout this study: decision-tree-like evaluation and a global tree evaluation scheme known as the rule of combination in the Dempster-Shafer Evidence Theory [5, 6]. A simple and fast method to get the class label of a given sample is to traverse the tree, starting from the root node to a leaf node, as in the decision tree approach. This approach does not provide fuzzy class memberships and therefore majority voting is the only possible scheme to combine the crisp decisions of the tree classifiers of the forest.

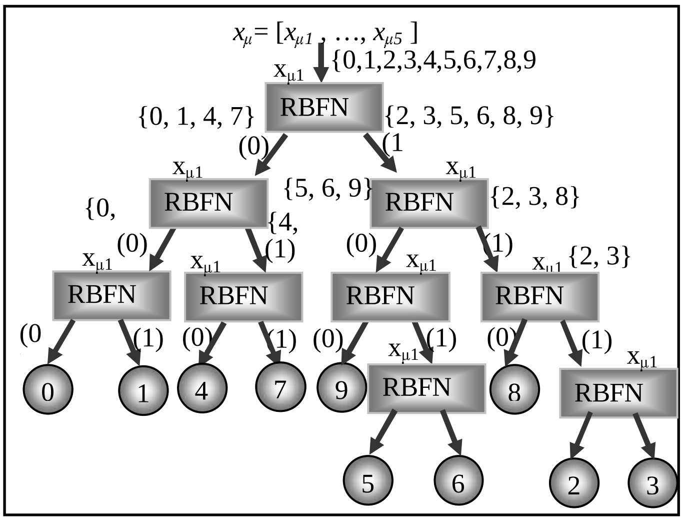

Fig. 3: Tree-Structured Classifier for handwritten digits using the image vector.

Dempster-Shafer evidence theory is a mathematical theory of evidence and is a tool for representing and combining evidences. The reasons for using this theory in the multiple classifiers combination, as discussed in [7], are: “the ability to easily represent evidences at different levels of abstraction and the possibility to combine evidences from different sources”. The Dempster-Shafer (DS) theory starts by assuming a universe of discourse consisting of a finite set of mutually exclusive atomic hypotheses Θ. Let 2^Θ denote the set of all subsets of Θ. Then a function m: $2^\Theta \rightarrow [0, 1]$ is called basic probability assignment (*bpa*) if it satisfies

$$m(\phi) = 0 \quad \text{and} \quad \sum_{A \subseteq \Theta} m(A) = 1 \tag{5}$$

It is possible to combine the basic probability assignments produced by *n* independent sources $m_1,\ldots, m_n$ using the orthogonal sum which is defined as

$$m(A) = K \sum_{\cap A_i = A} \prod_{1 \le i \le n} m_i(A_i) \tag{6}$$

where
$$K^{-1} = 1 - \sum_{\cap A_i = \phi} \prod_{1 \le i \le n} m_i(A_i) = \sum_{\cap A_i \ne \phi} \prod_{1 \le i \le n} m_i(A_i) \tag{7}$$

The normalization factor K indicates how much $m_1,\ldots, m_n$ are contradictory.

In our study, we assume that Θ is the set of class labels and the RBF networks outputs are transformed into *bpas* then the resulting *bpas* are combined using the orthogonal sum without normalization.

3 Application

The performance was evaluated using the handwritten STATLOG digits data set [4]. This data set consists of 10,000 images (1,000 images per class) and each image is represented by a 16x16 matrix containing the 8-bit grey values of each pixel (see Figure 4). Each sample is represented by five feature vectors as described in Table 1.

Feature	Description
image_vector	A 256-dim vector results from reshaping the 16x16 image matrix.
orienthisto	A 144-dim vector that represents 9 orientation histograms where an image matrix has been divided into 3x3 overlapped sub-images and a histogram is created for each sub-image.
pca_40	A feature vector results from projecting the *image_vector* onto the first 40 principal components of PCA.
rows_sum	A 160-dim vectors representing the sums over the rows of the original image and images results from rotating it 9 times.
cols_sum	A 160-dim vectors representing the columns over the rows of the original image and images results from rotating it 9 times.

Table 1. Description of the feature sets for the handwritten digits.

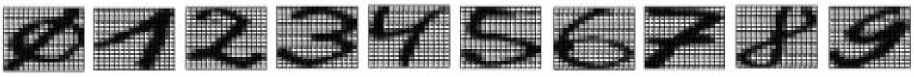

Fig. 4: A sample of the handwritten digits.

The RBF networks have been used as binary classifiers such that the hidden layer consists of 20 RBFs per class and the number of the input layer nodes equals to the dimension of the feature vector. The classification results are the average of one run of 10-fold cross-validation (CV). First, we construct a tree classifier for each possible view. Table 2 illustrates the performance of the five single-view tree classifiers using decision-tree-like method (DT) and Dempster-Shafer based method (DS) respectively.

TCM	*image_vector*	*orienthisto*	*pca_40*	*rows_sum*	*cols_sum*
DT	95.89%±0.47	96.05%±0.59	94.96%±0.81	94.07%±0.65	93.75%±0.95
DS	96.23%±0.55	96.51%±0.54	95.66%±0.61	94.52%±0.57	94.08%±0.97

Table 2. Results of the five Single-View Tree Classifiers for the handwritten digits.

Then, we construct two ensembles: one based on the 5 Single-View tree classifiers (3rd column in Table 3, MVF$_{\text{single}}$) and the other based on the 31 constructed classifiers (4th column in Table 3, MVF(31)). The results show that the performance of MVF$_{\text{single}}$ outperforms the best Single-View tree classifier and shows better performance than MVF(31). In addition, we found that some of the tree classifiers in MVF(31) are similar.

Ensemble Accuracy	TCM	FCM	MVF$_{\text{single}}$	MVF(31)	MVF(5)
	DT	MV	96.80%±0.44	94.08%±0.64	96.80%±0.44
	DS	MV	97.14%±0.45	94.59%±0.61	97.14%±0.45
		Min	97.43%±0.53	97.41%±0.52	97.43%±0.53
		Max	97.63%±0.54	**97.62%±0.51**	97.63%±0.54
		Mean	97.64%±0.57	95.69%±0.63	97.64%±0.57
		Prod	**97.71%±0.47**	96.51%±0.50	**97.71%±0.47**
Best Tree Accuracy			96.51%±0.54	96.62%±0.57	96.51%±0.54

Table 3. Results of the three Multi-View Forests for the handwritten digits.

3.1 Kappa Pruning

The Kappa pair-wise agreement measure [8] is used to find similar classifiers in order to remove them from the forest keeping only the 5 most diverse classifiers (the last column in Table 3, MVF(5)). Given two tree Classifiers that discriminate among L classes and m examples, the coincidence matrix element C_{ij} represents the number of examples that are assigned by the first classifier to class i and are assigned by the second classifier to class j. Then, the agreement measure κ is defined as follows:

$$\kappa = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2} \tag{8}$$

where $$\Theta_1 = \frac{\sum_{i=1}^{L} C_{ii}}{m} \quad \text{and} \quad \Theta_2 = \sum_{i=1}^{L}\left(\sum_{j=1}^{L}\frac{C_{ij}}{m}.\sum_{j=1}^{L}\frac{C_{ji}}{m}\right) \tag{9}$$

$\kappa = 1$ if the two classifiers agree on every example, and $\kappa = 0$ if the two classifiers agree with each other as we would expect from random agreements [8].

After applying the pruning method, it was found that the top five diverse classifiers are the Single-View tree classifiers. This means that the tree classifier based on *image_vector*, *orienthisto*, *pca_40*, *rows_sum* and *cols_sum* are diverse. Therefore, PCA shows its feasibility that it could contribute to increased diversity and accuracy.

4 Conclusion

In this study we have discussed a new ensemble creation method using multi-view tree-structured classifiers. The intended diversity in the proposed method will come from using different feature sets. Experiments demonstrate that Multi-view learning can improve the accuracy in complex problems with a large number of classes.

The fundamental assumption of multi-view learning, that each pattern must be represented with many independent feature sets, is not satisfied in many practical cases which negatively affect its applicability. Experiments show that the PCA-based feature set was independent from the original image feature vector. Therefore, the applicability of multi-view learning can be increased through applying different feature extraction algorithm such as PCA to create new feature sets.

Furthermore, Dempster-Shafer Evidence Theory based combination method was used to provide soft class labels. Therefore, a forest can be constructed not only by majority voting but also by minimum, maximum, mean, and product. The experiments show that soft combination methods outperform the crisp ones.

References

[1] L. I. Kuncheva and C. J. Whitaker, Ten measures of diversity in classifier ensembles: Limits for two classifiers. IEEE Workshop on Intelligent Sensor Processing, Birmingham, pages 1–10, 2001.

[2] F. Schwenker and G. Palm, Tree structured support vector machines for multi-class pattern recognition, *Multiple Classifier Systems*, MCS2001, pages 409-417, Springer-Verlag, 2001.

[3] M. Halkidi, Y. Batistakis and M. Vazirgiannis, Clustering Validity Checking Methods: Part II, *SIGMOD* Record 31(3): 19-27, 2002.

[4] F. Schwenker, H. Kestler and G. Palm, Three learning phases for radial-basis-function networks, *Neural Networks*, 14(4-5):439-458, Elsevier, 2001.

[5] A. P. Dempster, A generalization of bayesian inference, *Journal of the Royal Statistical Society*, 205–247, 1968.

[6] G. Shafer, A mathematical theory of evidence. (University Press, Princeton, 1976).

[7] R. Fay, F. Schwenker, C. Thiel and G. Palm, Hierarchical Neural Networks Utilising Dempster-Shafer Evidence Theory. In F. Schwenker and S. Marinai, editors, *proceedings of the 2nd IAPR workshop* (ANNPR 2006), Springer-Verlag LNAI 4087, pages 198-209, 2006.

[8] D. Margineantu and T. Dietterich, Pruning Adaptive Boosting, *proceedings of the 14th international conference on Machine Learning* (ICML 1997), pages 211-218, 1997.