ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

Classification of Bioacoustic Time Series by Training a Decision Template Fusion Mapping

F. Schwenker, C. Dietrich, G. Palm
University of Ulm, D-89069 Ulm, Germany

Abstract. Adaptable fusion techniques for the combination of local classifier decisions calculated from different feature subspaces are the topic of this paper. Decision template fusion is is discussed in the context of neural network learning algorithms, and applied to the recognition of bioacoustic time series.

1 Introduction

Combining the classification powers of several classifiers is regarded as a general problem in various pattern recognition applications [9, 4]. For the decison fusion *static* and *adaptable combining paradigms* [1, 5] have been proposed and discussed. In the static fusion mapping approach the individual classifiers of the ensemble are separetely trained by a single pass, and classifier fusion mapping is implemented by a predefined so-called *aggregation rule*. Adaptable fusion mappings are trained by a two-phase training procedure:

1. building the *Classifier Layer* consisting of a set of *first level classifiers* using training data $\mathcal{R}$, and
2. training the *Fusion Layer* performing a mapping of the classifier outputs (soft or crisp decisions) into the set of desired class labels by using a validation set $\mathcal{V}$.

The overall classifier architecture (see Figure 1) is a two-layered structure similar to *multilayer perceptrons* and *radial basis functions*.

2 Classification Fusion of Local Features

In this Section we propose a static and a trainable fusion architecture for time series classification. It is assumed that each time series is labeled with its corresponding class label $\omega \in \Omega$, $\Omega = \{1, ..., L\}$. A sliding window W^j covering a small part of the time series $s(t)_{t=1}^{T}$ is moved over the whole time series. For each window W^j, $j = 1, ..., \mathcal{J}$ a set of I features $\mathbf{x}_i(j) \in \mathbb{R}^{D_i}$, $i = 1, ..., I$ and $D_i \in \mathbb{N}$, is extracted. Typically $\mathcal{J}$, the number of time windows varies from

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

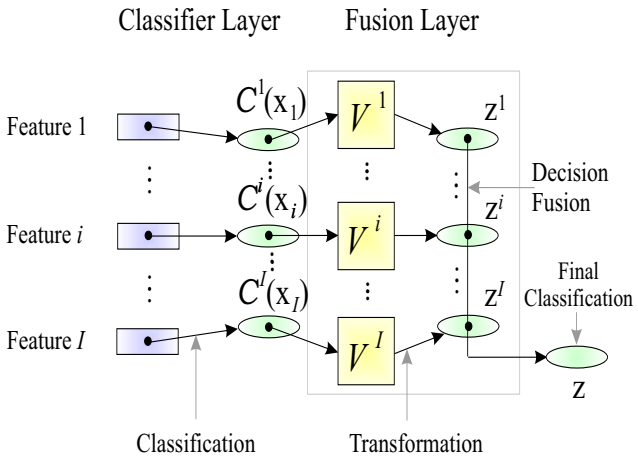

Figure 1: MCS with classifier layer and fusion layer. The combination of the classifier outputs $\mathcal{C}^i(\mathbf{x}_i)$, $i = 1, ..., I$ is accomplished through a fusion mapping $\mathcal{F}(\mathcal{C}^1(\mathbf{x}_1), ..., \mathcal{C}^I(\mathbf{x}_I))$. We consider fusion mappings where the classifier output $\mathcal{C}^i(\mathbf{x}_i)$ are linearly combined.

time series to time series. For $s(t)_{t=1}^T$ this leads to I feature streams $\mathbf{x}_i(j)_{j=1}^{\mathcal{J}}$. To determine the class membership of an input vector $\mathbf{x}$ to all L classes, a fuzzy-k-nearest-neighbour classifier is used [8]. Such a fuzzy classifier $\mathcal{C}$ is defined as a mapping $\mathcal{C} : \mathbb{R}^D \rightarrow [0,1]^L$, i.e., the output $\mathcal{C}(\mathbf{x}) = (\mathcal{C}_1(\mathbf{x}), ..., \mathcal{C}_L(\mathbf{x}))$ contains the memberships of $\mathbf{x}$ to each class. After normalization by $\mathcal{C}_l(\mathbf{x}) := \delta_l(\mathbf{x}) / \sum_{l=1}^L \delta_l(\mathbf{x})$ we get *soft labels* $\mathcal{C}(\mathbf{x}) \in \Delta$ with

$$\Delta := \{(\mathbf{y}_1, ..., \mathbf{y}_L) \in [0,1]^L | \sum_{l=1}^{L} \mathbf{y}_l = 1\} \tag{1}$$

2.1 CDT-Architecture

The CDT fusion performs the classification in three steps:

1.) Classification of single feature vectors (C-step)
For each feature $i = 1, ..., I$ a classifier $\mathcal{C}^i$ is given through a mapping

$$\mathcal{C}^i : \mathbb{R}^{D_i} \rightarrow \Delta. \tag{2}$$

Thus, for each time window W^j, $j = 1, ..., \mathcal{J}$, I classification results $\mathcal{C}^1(\mathbf{x}_1(j)), ..., \mathcal{C}^I(\mathbf{x}_I(j))$ based on the individual features $\mathbf{x}_1(j), ..., \mathbf{x}_I(j)$ are calculated.

2.) Decision fusion of the local decisions (D-step)
For each time window W^j the I classification results are combined into

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

a local decision $\mathbf{z}^j \in \Delta$ through a fusion mapping $\mathcal{F} : \Delta^I \rightarrow \Delta$

$$\mathbf{z}^j := \mathcal{F}(\mathcal{C}^1(\mathbf{x}_1(j)), ..., \mathcal{C}^I(\mathbf{x}_I(j))), \quad j = 1, ..., \mathcal{J}. \tag{3}$$

3.) Temporal fusion of decisions over the whole time series (T-step)
The combination of the local decisions of the whole set of time windows W^j, $j = 1, ..., \mathcal{J}$ is given through

$$\mathbf{z}^o := \mathcal{F}(\mathbf{z}^1, ..., \mathbf{z}^{\mathcal{J}}). \tag{4}$$

In the numerical experiments we applied averaging : $\mathcal{F}(\mathbf{z}^1, ..., \mathbf{z}^N) = \sum_{n=1}^{N} \mathbf{z}^n$ for decision fusion (see Eq. 3) and the temporal integration (see Eq 4).

2.2 Decision Templates

The concept of decision templates as a trainable aggregation rule was introduced by KUNCHEVA [5, 6]. For a trained classifier $\mathcal{C}$, *decision templates* $\mathcal{T}^\omega$ for each class $\omega \in \Omega$ can be calculated by the average of the local classifier outputs $\mathcal{C}(\mathbf{x}^\mu(\cdot))$ for inputs $\mathbf{x}^\mu(\cdot)$ from a class ω [6]:

$$\mathcal{T}^\omega := \frac{1}{|\mathcal{V}^\omega|} \sum_{\mathbf{x}^\mu(j) \in \mathcal{V}^\omega} \mathcal{C}(\mathbf{x}^\mu) \tag{5}$$

$\mathcal{V}^\omega$ is a validation set of $\mathbb{R}^D \times \{\omega\}$ differnent from the classifier training set. Decision template $\mathcal{T}^\omega \in \Delta$ can be interpreted as a characteristic classifier output for the inputs $\mathbf{x}^\mu(\cdot)$ of $\mathcal{V}^\omega$. In the case of I input features with classifier mappings $\mathcal{C}^i : \mathbb{R}^{D_i} \rightarrow \Delta$, $i = 1, ..., I$ the *decision template* $\mathcal{T}^\omega$ of class ω is given by a $(I \times L)$-matrix

$$\mathcal{T}^\omega := \left(\mathcal{T}_1^\omega, \ldots, \mathcal{T}_I^\omega\right) \in \Delta^I. \tag{6}$$

Hereby $\mathcal{T}_i^\omega \in \Delta$ is the decision template of the i-th feature space $\mathbb{R}^{D_i}$ and target class ω. The local decision profile $\mathcal{P}^j$ for an input $X(j) = (\mathbf{x}_1(j), ..., \mathbf{x}_I(j))$ is given by the individual classifier outputs of the I classifiers

$$\mathcal{P}^j(X(j)) = [\mathcal{C}^1(\mathbf{x}_1(j)), ..., \mathcal{C}^{\mathcal{J}}(\mathbf{x}_{\mathcal{J}}(j))]^T \in \Delta^I. \tag{7}$$

Classifiers $\mathcal{C}^1, ..., \mathcal{C}^I$ are applied to calculate the local decision profile $\mathcal{P}^j$ (see step (a) in Algorithm DT). Then for each class $\omega \in \Omega$ a local class membership value $\mathbf{z}_\omega^j$ based on a similarity measure $\mathcal{S}$ between the decision profile $\mathcal{P}^j$ and the decision template $\mathcal{T}^\omega$ is calculated (see step (b) and Eq. 8). After temporal integration of local decisions (see step (c)) the class with the maximum membership ω^* is the final decision, see step (d). As similarity measure the normalized Euclidean distance was used:

$$\mathcal{S}(\mathcal{P}, \mathcal{T}^\omega) := 1 - \frac{1}{2I} \sum_{i=1}^{I} ||\mathcal{P}_{i,\cdot} - \mathcal{T}_{i,\cdot}^\omega|| \quad \in [0, 1]. \tag{8}$$

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

Algorithm $\omega^* = \mathrm{DT}((X(j))_{j=1}^{\mathcal{J}}, (\mathcal{T}^\omega)_{\omega=1}^{L})$
  foreach $j = 1, ..., \mathcal{J}$
(a) $\mathcal{P}^j = [\mathcal{C}^1(\mathbf{x}_1(j)), ..., \mathcal{C}^I(\mathbf{x}_I(j))]^T$
    foreach $\omega \in \Omega$
(b) $\mathbf{z}_\omega^j = \mathcal{S}(\mathcal{P}^j, \mathcal{T}^\omega)$
    end
  end
(c) $\mathbf{z} = \mathcal{F}(\mathbf{z}^1, ..., \mathbf{z}^{\mathcal{J}})$
(d) $\omega^* = \underset{\omega}{\mathrm{argmax}}(\mathbf{z}_\omega)$

Algorithm **DT**: Classification of time series $(X(j))_{j=1}^{\mathcal{J}}$ consisting of $\mathcal{J}$ local feature vectors with decision templates.

3 Decision Templates and Neural Networks

The training of a multiple classifier system (MCS) with a decision template layer is very similar to radial basis function (RBF) network training, where the network is learnt by two or three learning phases. In Fig. 1 a two-layer multiple classifier system (MCS) is presented. Here the combination of the classifier outputs $\mathcal{C}^i(\mathbf{x}_i)$, $i = 1, ..., I$ is accomplished through a fusion mapping $\mathcal{F}(\mathcal{C}^1(\mathbf{x}_1), ..., \mathcal{C}^I(\mathbf{x}_I))$, where the classifier outputs $\mathcal{C}^i(\mathbf{x}_i)$ are multiplied by matrices V^i, $i = 1, ..., I$ and the decisions $\mathbf{z}^1, ..., \mathbf{z}^I$ are combined by averaging. The matrices V^i are adapted through an supervised learning phase. For this, it is assumed that the desired classifier outputs $\omega^\mu \in \Omega$ of inputs $\mathbf{x}_i^\mu \in \mathcal{V}$ are given by the $(L \times M)$-matrix Y (L the number of classes, M the number of training patterns) defined by the 1-out-of-L encoding scheme for class labels $Y_{l,\mu} = 1$ iff $l = \omega^\mu$. Corresponding to C_i the μ-th column $Y_{\cdot,\mu} \in \Delta$ contains the binary coded target output of feature vector $\mathbf{x}_i^\mu \in \mathcal{V}$.

Now, let $\mathcal{T}_i^\omega \in \Delta$ be the decision template of the i-th feature space and for class ω (see Eq. 5). Then for each classifier $i = 1, ..., I$, a $(L \times L)$-decision template V^i is given by the decision templates of the i-th feature space

$$V^i := \left(\mathcal{T}_i^1, \ldots, \mathcal{T}_i^L\right) \in \Delta^L. \tag{9}$$

this can be written as

$$V^i := (YY^{\mathrm{T}})^{-1} \underbrace{(YC_i^{\mathrm{T}})}_{=:W^i}. \tag{10}$$

So in the decision template approach the linear mappings V^i are basically given by the the confusion matrix $W^i := YC_i^{\mathrm{T}}$.

A second approach to calculate a linear decision fusion mapping $\mathcal{F}$ is given by the minimal least squares solution between the classifier outputs and the target outputs. This error function is often used to train single layer neural

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

networks, e.g. the output layer of RBF networks. The solution of the error function can be expressed by

$$V^i := \lim_{\alpha \to 0_+} \underbrace{(YC_i^{\mathrm{T}})}_{=W^i} (C_i C_i^{\mathrm{T}} + \alpha I)^{-1}. \tag{11}$$

Thus, if $C_i C_i^{\mathrm{T}}$ is regular the solution of V^i reduce to

$$V^i = W_i (C_i C_i^{\mathrm{T}})^{-1}. \tag{12}$$

As for the the decision template approach, the matrices $V^1, ..., V^I$ based on the confusion matrices $W^i := YC_i^{\mathrm{T}}$.

feature comp.	average	DT
I	29.35	20.20
II	22.64	18.91
III	26.87	22.49

Table 1: Error rates(in %) for the CDT architecture with averaging fusion mapping $\mathcal{F}$ and the decision template fusion; results for three different feature subspace combinations are given.

4 Application and Conclusion

We present results achieved by testing the algorithms on a dataset which contains sound patterns from 22 different katydid species. The dataset contains recordings of 6 to 12 different individuals per species. Recordings are provided from Ingrisch [3], Nischk [7] and Heller [2]. Sound patterns are stored in the WAV-format (sampling rate ranges fro 44.1 kHz to 500.000 kHz, 16 Bit sampling accuracy). The katydid songs consist of sequences of sound patterns called syllables. Based on these syllables (sequences of so-called impulses) the katydid species are classified [7]. Therefore we determine the on- and off-sets of these single impulses in order to get the relevant parts of the signal. The time windows W^t, $t = 1, ..., T$ are aligned at the onsets and offsets of the pulses and the features: pulse length, pulse distances and pulse frequencies, time encoded signals, and energy contours of pulses are calculated inside these windows (details about the feature extraction can be found in [1]). In Table 1 the classification results for the proposed fusion schemes are given for three feature compositions. In all cases the decision template approach outperform the static averaging fusion mapping.

The basic conclusion of our experiments is that the decision template approach can improve the classifier performance in comparison to the combination of multiple classifiers with averaging. Finally, and independent from these

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 555-560

questions of fundamental research, the classifier system described here could be implemented in order to detect, classify and monitor biodiversity. Additionally we have shown that the solution in decision templates approach of MCS is similar to the least squares solution for linear single layer neural networks.

References

[1] C. Dietrich, F. Schwenker, and G. Palm. Classification of time series utilizing temporal and decision fusion. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 378–387. Springer, 2001.

[2] K. G. Heller. *Bioakustik der Europäischen Laubheuschrecken. Ökologie in Forschung und Anwendung.* Margraf, Weikersheim, 1988.

[3] S. Ingrisch. Taxonomy, stridulation and development of Podoscirtinae from Thailand. *Senckenbergiana biologica*, 77:47–75, 1997.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE PAMI*, 20(3):226–239, 1998.

[5] L. I. Kuncheva. Using measures of similarity and inclusion for multiple classifier fusion by decision templates. *Fuzzy Sets and Systems*, 122(3):401–407, 2001.

[6] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. Decision templates for multiple classifier fusion. *Pattern Recognition*, 34(2):299–314, 2001.

[7] F. Nischk. *Die Grillengesellschaften zweier neotropischer Waldökosysteme in Ecuador.* PhD thesis, University of Köln, Germany, 1999.

[8] S. Singh. 2D spiral pattern recognition with possibilistic measures. *Pattern Recognition Letters*, 19(2):141–147, 1998.

[9] Lei Xu, Adam Krzyzak, and Ching Y. Suen. Methods of combining multiple classifiers and their applications in handwritten character recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.