Combination of Measurement-Level Classifiers: Output Normalization by Dynamic Time Warping

G. Pirlo[(*)(^)], *D. Impedovo*[(§)(^)], *C.A.Trullo*[(*)(^)], *E. Stasolla*[(*)(^)]

[(*)] *Dipartimento di Informatica– Università degli Studi di Bari - via Orabona 4, 70126 – Bari*
[(§)] *Dipartimento di Elettrotecnica ed Elettronica-Politecnico di Bari –Via Orabona 4, 70126 – Bari*
[(^)] *Centro "Rete Puglia" – Università degli Studi di Bari – via G. Petroni 15/F.1- 70100 Bari*
pirlo@di.uniba.it

Abstract

Classifier combination is a powerful strategy to support useful solutions in difficult classification problems. Notwithstanding, the effectiveness of a multi-classifier system strongly depends on the decision fusion strategies. In this field, one of the most significant aspects concerns output normalization, when classifiers decisions are provided at measurement level.

This paper presents a new approach for output normalization that uses Dynamic Time Warping (DTW). Some experimental tests have been carried out in the field of handwritten digit recognition. The proposed approach is superior to other output normalization algorithms in the literature.

1. Introduction

In the last years, classifier combination has been demonstrated to be an effective strategy to solve difficult classification problems, like those related to on-line and off-line handwriting recognition [1,2].

When classifier combination is considered, the final classification decision is obtained combining the decisions of the individual classifiers. On the basis of the kind of decisions combined, the methods for classifier combination can be divided into three categories: - In the *Measurement-level*, combination methods combine values provided by individual classifiers as a measure of the degree of membership of the input pattern to each class; - In *Ranked-level*, combination methods combine ranked lists of class labels ordered according to the degree of membership of the input pattern; - In *Abstract-level*, combination methods combine simple class labels [2, 3].

Among the three categories, the combination of classifiers at the measurement-level is expected to be the most effective, since it uses all information available. Unfortunately, the performance of a multi-classifier system based on measurement-level classifiers, strongly depends on the effectiveness of the strategy used for normalizing outputs of the individual classifiers. Indeed, the output produced by each individual classifiers can depend on several aspects: the feature used, the classification strategy, the similarity or dissimilarity measure that has been used and so on. The result is that different classifiers produce outputs in different ranges, therefore they are difficult to compare. In addition, also when the output range is the same for two different classifiers, the meaning of an output value produced by the two classifiers can be different, when – for instance – it is obtained by different similarity/dissimilarity measures or membership functions. Therefore, incomparability of the classifier output scores is a major problem in the combination of different classifiers. In order to deal with this problem, the measurement level classifier outputs are generally normalized [4,5].

In the past, several approaches have been considered for output normalization based on output range resizing, statistical adaptation, characteristic functions [6,7,8].

This paper presents a new technique that performs output normalization by Dynamic Time Warping and shows its superiority with respect to the approaches in the literature.

The organization of the paper is as follows: Section 2 presents the state-of-the-art techniques for output normalization. The new technique is introduced in Section 3. Section 4 presents the experimental results, carried out in the field of handwritten numeral recognition, achieved by applying the different output normalization techniques to a well-defined multiexpert system. The conclusions of the present work and some directions for further research are highlighted in Section 5.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.74

2. Techniques for Output Normalization

Let A be a measurement–level classifier that associates input patterns to the classes of the set $\Omega = \{\omega_1, \omega_2, \ldots. \omega_i, \ldots, \omega_M\}$. In this case given an unknown pattern x, A produces a set of output values $A(x)=(a_1, a_2, \ldots, a_i, \ldots, a_M) = (A(x, \omega_1), A(x, \omega_2), \ldots, A(x, \omega_i), \ldots, A(x, \omega_M))$ where each matching score $a_i = A(x, \omega_i)$ is a confidence value supporting the evidence that x belongs to ω_i. In the ideal case a_i is the a-posteriori probability of ω_i given the input pattern x. Sometimes, a_i is just an approximation of the a-posteriori probability. In many practical cases, however, a_i is the result of a matching between the input pattern x and the class ω_i, computed using similarity or dissimilarity measures. Therefore, scales and distributions of the output values provided by different classifiers are generally very diverse, since they strongly depend on many parameters, such as features used for pattern matching, similarity/dissimilarity measures, etc. Therefore, the problem of output normalization is very significant when results from different classifiers have to be combined, in order to exploit potential of classifier combination.

In this Section three different techniques in the literature for output normalization will be briefly described: the min-max technique, the z-score technique and the normalization technique based on the characteristic function.

2.1 Min-Max

Min-max normalization [6] is the simplest normalization technique that is best-suited for the cases where the bounds of the scores produced by a classifier are known:

- S_{min}= minimum of the scores;
- S_{max}= maximum of the scores.

In this case, given a set of matching scores a_i, i=1,2,…, M, the set of normalized scores is given by:

$$a'_i = \frac{a'_i - S_{\min}}{S_{\max} - S_{\min}}. \quad (1)$$

Of course, when the matching score are not bounded, S_{min} and S_{max} can be estimated.

2.2 Z-score

Z-score normalization is the most commonly used score normalization technique [6]. This technique works well if prior knowledge about the average score and the score variations of the classifier is available. If there is no a priori knowledge about the nature of the matching algorithm, the mean and the standard deviation of the scores should be estimated from a given set of matching scores. Given a set of matching scores a_i, i=1,2,…, M, the set of normalized scores is given by:

$$a'_i = \frac{a'i - \mu}{\sigma}. \quad (2)$$

where μ is the arithmetic mean and σ is the standard deviation of the given data set.

2.3 Characteristic function

Output normalization can be achieved by aligning the accumulated recognition rate, that is a continuous, monotone growing function over classifier output [7,8]. For this purpose the correct recognition function of a classifier is first computed, that counts, for each likelihood value, the number of correctly recognized samples $n_{correct}(j)$ [7,8]. In this case, given a set of matching scores a_i, i=1,2,…, M, the set of normalized scores is given by:

$$a'_i = a_i + charf(a_i) \quad (3)$$

with the function *charf(·)* is the characteristic function of a classifier and it is defined as [7]:

$$charf(a_i) = a_{max} * r_i - a_i \quad (4)$$

$$r_i = \frac{\sum_{j=1}^{i} n_{correct}(j)}{N} \quad (5)$$

where:

- a_{max} is the maximum possible output of a classifier;
- N is the number of overall patterns;
- r_i is the partially accumulated recognition rate.

Of course, to compute the characteristic function of a classifier, an extra sample set is necessary, which should be independent from the training set.

3. Output Normalization by DTW

In this Section a new technique is proposed for normalizing the output of classifiers. It uses a Dynamic Time Warping (DTW) technique to match the cumulative recognition rate of a classifier against a standard curve, that is the same for all classifiers. More precisely, in this paper, the accumulated Gaussian probability function has been considered as standard curve ($\mu = 0$, $\sigma = 1$). Now, let

$$R = (r_1, r_2, \ldots\ldots, r_L) \quad (6)$$

be the L-point discrete version of the accumulated recognition function and

$$G = (g_1, g_2, \ldots\ldots, g_L) \quad (7)$$

be the L-point discrete version of the accumulated probability function of the standard Gaussian distribution. A warping function between R and G is any sequence of couples of indexes identifying points of R and G to be joined [9]:

$$W(R,G)=c_1,c_2,\ldots,c_T, \quad (8)$$

where $c_t=(i_t,j_t)$, being T,i_t,j_t integers, $1\le t\le T$, $1\le i_t\le L$, $1\le j_t\le L$.

Now, if a distance measure is considered between points of R and G,

$$d(c_t) = d(r_t, g_t), \quad (9)$$

we can associate to W(R, G) the dissimilarity measure

$$D_{W(R,G)} = \sum_{t=1}^{T} d(c_t). \quad (10)$$

The elastic matching procedure detects the optimal warping function $W^*(R, G)= c^*_1,c^*_2,\ldots,c^*_{T^*}$ which satisfies the monotonicity, continuity and boundary conditions, and for which it results [9]:

$$D_{W^*(R,G)} = \min_{W(R,G)} D_{W(R,G)}. \quad (11)$$

Now let be W*(R, G) the optimal warping function for the matching between R and G, given a matching score a_i, i=1,2,…,n, its normalized value is given by:

$$a'_i = \frac{1}{card(G_i)} \cdot \sum_{g_q \in G_i} g_q \quad (12)$$

where G_i is the set of samples of G associated to the sample r_i of the accumulated recognition function, i.e.:

$$G_i = \{ g_q \mid (q,i) \in W^*(R, G) \}. \quad (13)$$

In other words, the normalized value of a score a_i is given by the average value of the points of the accumulated Gaussian function which corresponds to the sample r_i, being r_i the sample of the accumulated recognition function that corresponds to a_i. Thus, Dynamic Time Warping is used to map the value of the accumulated recognition function r_i (corresponding to the confidence value a_i) on the standard curve (the accumulated Gaussian function). The average of the values on the standard curve that corresponds to r_i is considered as the normalized value a'_i corresponding to a_i.

4. Experimental Results

In order to test the proposed technique the multiclassifier system in Figure 1 has been considered. The input pattern x is fed to K individual classifiers in parallel. Each classifier A_i provides its response $A_k(x)$, k=1,2.,..K. The responses obtained by all the classifiers are then combined, according to a suitable combination strategy, to obtain the final results $E(x)= E(A_1(x), A_2(x),\ldots, A_K(x))$ [1, 2].

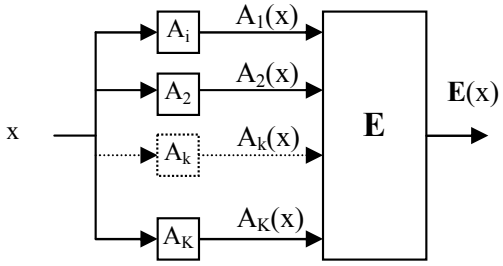

Figure. 1: Multiclassifier System

The system has been designed for handwritten digits recognition and four different classifiers have been considered (K=4). The classifiers use:

A_1: Histogram-based classifier (Fig. 2a);

A_2: Intersection-based classifier (Fig. 2b);

A_3: Zoning-based classifier (Fig. 2c);

A_4: Pattern Matching classifier (Fig. 2d).

The description of the classifiers is beyond the aims of this paper and the interested reader can found more details in refs. [10, 11].

A simple rule [12] has been considered for output combination. According to this rule let x an unknown input pattern and a_i^k the confidence value confidence value, produced by the classifier A_k, supporting the fact that x belongs to ω_i. Now, let be

$$B_i = \sum_{s=1}^{K} a'^{s}_{i}, \quad i=1,2,\ldots,M \quad (14)$$

with a'^{k}_{i} being the normalized version of a_i^k, the pattern x is classified to the class ω_j if and only if

$$B_j = \max_{(i=1,2,\ldots K)} B_i \quad (15a)$$

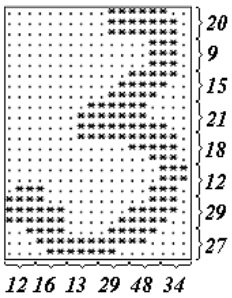

Figure 2.a

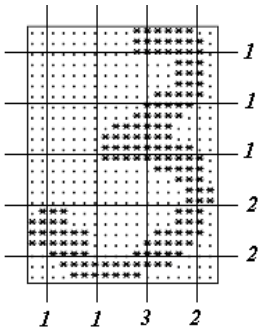

Figure 2.b

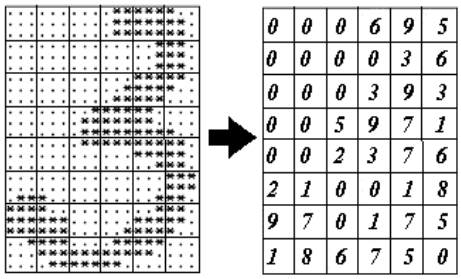

Figure 2.c

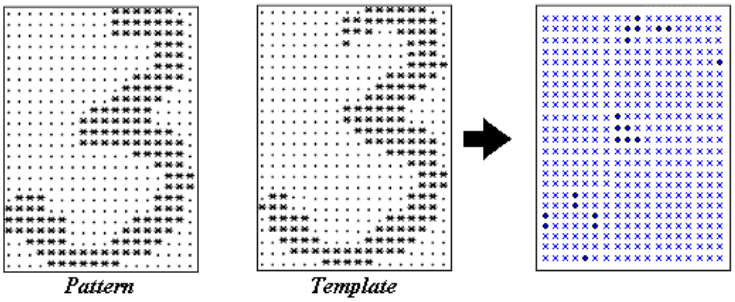

Figure 2.d

and

$$| B_j - \max_{(i=1,2,\ldots K \text{ and } i\neq j)} B_i | \; / \; B_j \leq \varepsilon \qquad (15b)$$

being ε a threshold value.

Furthermore, the CEDAR database has been considered [13]: 18467 patterns for learning (BR directory) and 2189 patterns for the test (BS directory).

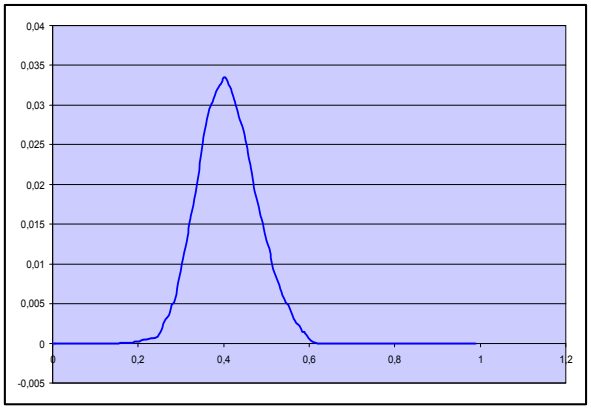

Figure 3. Correct Recognition Function

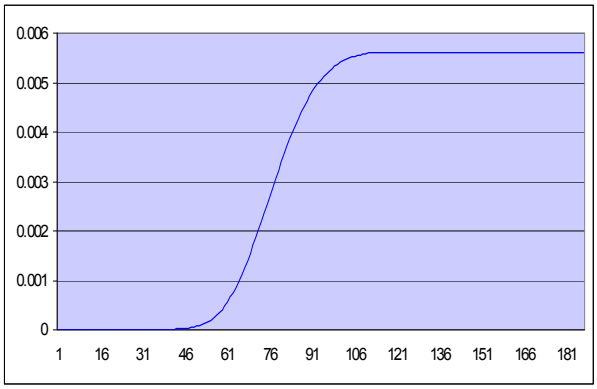

Figure 4. Accumulated Recognition Function

Figure 3 shows correct recognition function for the output of the histogram-based classifier, that is the distribution of scores identifying recognized patterns. Figure 4 shows the accumulated recognition function for the output of the histogram-based classifier. Figure 5 shows the results of DTW between the accumulated recognition functions of the histogram-based classifier (lower function) and the accumulated Gaussian probability function (upper function). In this case the Euclidean distance is used for matching.

Tables 1 and 2 report the performance of the multiexpert system, depending on the normalization technique.

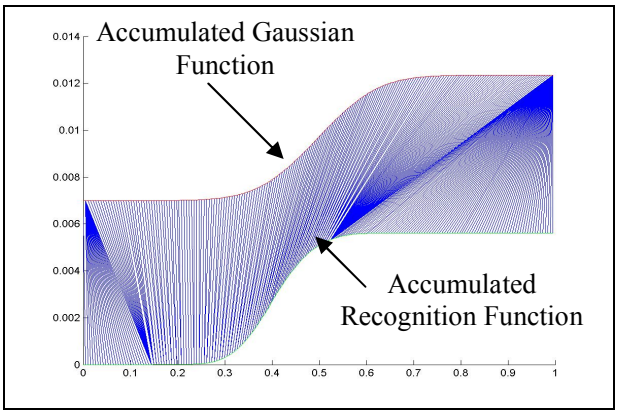

Figure 5. Warping between accumulated recognition function (lower curve) and accumulated Gaussian function (upper curve).

Table 1. Performance Analysis (ε=0%)

Normalization	Recog.	Reliab.
None	80,1%	80%
MIN-MAX	79,2%	79%
Z-Score	72,6%	72%
Characteristic Function	76,5%	76%
DTW-based	80,5%	80%

Table 2. Performance Analysis (ε=5%)

Normalization	Recog.	Reliab.
None	23%	100%
MIN-MAX	50,4%	94%
Z-Score	54%	87%
Characteristic Function	75,3%	77%
DTW-based	75,7%	85%

Table 1 reports the recognition rate and the reliability rate at zero rejects (ε=0). The improvement of the DTW-based technique, in terms of recognition rate, is equal to 0,4%, 1.6%, 10.8%, 5.2%, with respect to the case of no normalization, to the MIN-MAX, Z-Score, Characteristic Function techniques, respectively. In terms of reliability rate the improvement is respectively of 0%, 1%, 11%, 1.4%.

Table 2 reports the recognition rate and the reliability rate for ε=5%. In this case, only the new technique and the Characteristic Function technique provides acceptable results. In fact the recognition rates are absolutely unacceptable when no normalization is performed or when the MIN-MAX and Z-Score techniques are applied. Now, when the recognition rate is considered, the improvement of the DTW-based technique is equal to 0.5% with respect to the Characteristic Function technique. Conversely, when the reliability rate is considered, the improvement of the DTW-based technique is equal to 10.3%.

5. Conclusion

In this paper a new technique is presented for output normalization of measurement-level classifiers.

The approach used DTW for matching the accumulated recognition function of the classifiers on a standard accumulated function obtained from a Gaussian distribution.

The experimental results, obtained by a multiexpert system for handwritten numeral recognition, demonstrate the superiority of the proposed technique with respect to other approaches in the literature.

References

[1] R. Plamondon and S. N.Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Trans. PAMI*, vol.22, no. 1, Jan. 2000.

[2] L. Xu, A. Krzyzak, C. Y. Suen, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition" , *IEEE Transaction on Systems, Man and Cybernetics*- Vol. 22, N. 3, 1992, pp. 418-435.

[3] F. Kimura , M. Shridhar, "Handwritten Numerical Recognition Based on Multiple Algorithms", *Pattern Recognition*, Vol. 24, n. 10, pp. 969-983, 1991.

[4] H. Altinçay and . Demirekler, "Undesirable effects of output normalization in multiple classifier systems", *Pattern Recognition Letters*, Vol.24 , No. 9-10, 2003, pp. 1163 – 1170.

[5] C.-L. Liu, "Classifier combination based on confidence transformation", *Pattern Recognition,*Vol. 38, No. 1, January 2005, Pages 11-28

[6] A. Jain, K. Nandakumar, A. Ross, "Score normalization in multimodal biometric systems", *Pattern Recogniton*, Vol. 38, 2005, pp. 2270-2285.

[7] O. Velek, S. Jaeger, M. Nakagawa, "A New Warping Technique for Normalizing Likelihood of Multiple Classifiers and Effectiveness in Combined On-Line/Off-Line Japanese Character Recognition", *Proc.* IWFHR 2002,Niagara-on-the-lake,Canada,pp.177-182.

[8] O. Velek, S. Jaeger, M. Nakagawa, "Accumulated-Recognition-Rate Normalization for Combining Multiple On-Line/Off-Line Japanese Character Classifiers Tested on a Large Database", *Proc.* MCS 2003, Guildford, UK.

[9] L.R.Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Inc., 1993.

[10] G. Dimauro, S. Impedovo, G. Pirlo, A. Salzo, "Automatic Bankchecks Processing: A New Engineered System", *IJDAR*, Vol.11, N.4, World Scientific Publ., Singapore, 1997, pp.1-38.

[11] O.D.Trier, A.K.Jain , T.Taxt, "Feature Extraction Methods For Character Recognition – A Survey", *Pattern Recognition*, Vol. 29, n.4, pp. 641-662, 1996.

[12] J. Kittler, M. Hatef, R.P.W. Duin, J. Matias, "On combining classifiers", *IEEE Trans. on PAMI*, Vol.20, no.3, pp.226-239, 1998.

[13] J. Hull, "A database for handwritten text recognition research", *IEEE Trans. on PAMI*, Vol. 16, n. 5, pp. 550–554, 1994.