Multilayer Perceptrons Combination Applied To Handwritten Character Recognition

Bernard GOSSELIN
Circuit Theory & Signal Processing Lab, Faculté Polytechnique de Mons
Bd Dolez 31, B-7000 Mons, BELGIUM

Abstract: Several methods of combination of Multilayer Perceptrons (MLPs) for handwritten character recognition are presented and discussed. Recognition tests have shown that cooperation of neural networks using different features vectors can reduce significantly the overall misclassification error rate. Additionally, the MLPs that are combined are the results of the experiments that were previously performed in order to optimize the recognition process when using a single MLP. So, all the combination methods that are proposed are very easy to carry out. The final recognition system consists of a cascade association of small MLPs, which allows minimization of the overall recognition time while retaining a high recognition rate. This system appears to be 2.5 times faster than the best of the individual MLPs, while offering a recognition rate of 99.8 % on unconstrained digits extracted from the NIST 3 database.

Keywords: Optical Character Recognition, Features Extraction, Multilayer Perceptron.

1 Introduction

Artificial Neural Networks, and especially Multilayer Perceptrons, have shown good capabilities in performing handwritten character recognition. However, their performance is strongly affected by the quality of the representation of the characters. This may require a large number of parameters to represent the character, which then results in difficulty in establishing the rules for recognition. In other words the MLPs become difficult to train. Moreover, the greater the size of the network, the greater is the computation time. This can greatly restrict their practical use. So, it is necessary to perform efficient features extraction on the one hand, and to optimize the lay-out of the artificial neural network on the other hand.

This paper is organized as follows: section 2 presents two kinds of features extractors that can be used for off-line handwritten character recognition. Section 3 proposes a way to reduce the dimension of representation by selecting only the features that provide a high discrimination ability. Section 4 sets out the recognition results that were obtained for each kind of features vector when using a single MLP to classify handwritten digits extracted from the NIST Special Database 3 [1]. Section 5 shows that, as the misclassification errors of the two MLPs don't occur in the same samples, it is possible to reduce the overall error rate by combining the two artificial neural networks. However, the systematic use of the two MLPs requires a significant increase in the computation time. Sections 6 and 7 then present several methods of cascade combination of the neural networks, which are based on the rejection of characters by a MLP when the level of its outputs are below a pre-defined threshold. Section 8 summarizes the results of this work.

2 Features Extraction

Numerous methods of features extraction for character recognition have been developed [2-6]. As they are often too dependent on the particular problem for which they have been developed, not all of them allow an accurate recognition of unconstrained handwritten characters. However, two of these methods have been shown to be particularly efficient [7], and are described below.

2.1 The Averaged Pixels method

This method is the easiest one to make use of, and consists in reducing the size of each character to a normalized dimension. A grid area of 16 by 16 is superimposed on the image of the character, and the averaged value of the pixels of each area is computed. In order to keep information about the original aspect of the character, the ratio between the initial width and height of the character is also computed and included in the feature vector, which then consists of 257 components.

2.2 The Normalized Contour Analysis method

A contour analysis is performed on the normalized characters obtained by use of the first method. It consists of sending probes up to the character from several directions. The length of each probe is the ordinate, according to the search direction, of the first nonbackground pixel met. In order to get scale invariance and normalized values between 0 and 1, the length of each probe is divided by its highest possible value. Eight directions from the outside of the matrix to the inside and two directions from the inside to the outside are used. In addition, the number of intersections between the layout of the character and vertical and horizontal lines is also taken into account. This produces a feature vector consisting of 193 components, including the original aspect ratio of the character.

3 Features Selection

Using the Averaged Pixels method or the Normalized Contour Analysis method, the number of features remains too high and needs to be restricted. First of all, so as to eliminate redundancy, a Principal Component Analysis is performed. As each component of the features vector brings its own information, its ability for discrimination can be estimated thanks to the use of the generalized Fisher's criteria [8]:

$$D = \frac{\overline{\mu^2} - \mu^2}{\overline{\sigma^2}}$$

Where μ^2 is the square of the mean value of the considered feature computed over all classes, and $\overline{\mu^2}$ and $\overline{\sigma^2}$ are respectively the averaged values of the square of its mean value and of its variance computed over each class.

The higher this ratio is, the higher is the discrimination ability of the concerned feature. So, only the features for which this ratio is greater than a given threshold are retained.

In order to give the same *a priori* importance to each feature, the Principal Components Analysis should be normed. In this case, the resulting features have a mean value equal to 0, and a variance equal to 1. So, the expression for the discrimination ability becomes very simple, leading to an easy use of it:

$$D = \frac{\overline{\mu^2}}{1 - \overline{\mu^2}}$$

4 Results

Tests have been performed on handwritten digits extracted from the NIST Special Database 3 [1]. The training set contained 5000 samples of each class, and the test set, another 2000. The two methods of features extraction were applied. For each of them, a selection of discriminant features was carried out as described in part 3. For both features vectors, more than the half of their components showed a very poor discrimination ability. All these features could then be considered to bring more noise than useful information, and so, they were removed. The discrimination ability of the remaining features evolved in stages, so that the number of attempts that were required to be performed in order to find the most convenient number of features was very limited.

A multilayer perceptron with one hidden layer was trained to classify the characters. The use of the first features extractor (AP) led to a recognition rate of 97.8 % on the test set, with only 46 features selected from 257. The second features extractor (NCA) allowed a recognition rate of 98.7 % to be reached, with 69 components selected from 193. Although there are more selected components in the second case, the second network contains only 60 hidden units, whereas the first one contains twice as many. So, the recognition process in the second case is 55 % faster.

The recognition rates that have been obtained show the accuracy of the method of features selection. As the notion of rejection of characters is introduced when the value of the outputs of a MLP are below a given threshold, the evolution of the error rate according to the rejection rate can be measured (Figure 1).

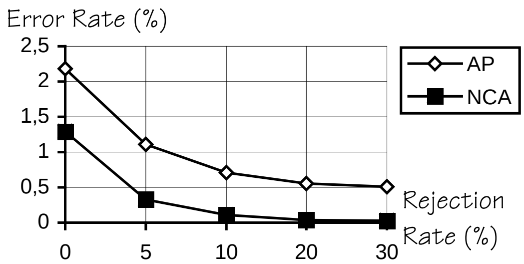

Figure 1: Evolution of the Error Rate according to the Rejection Rate

5 Parallel Cooperation of Multilayer Perceptrons

While one of the MLPs is more efficient, in terms of recognition performance as well as of computation time, it appeared that the misclassification errors of the two neural networks did not occur for the same samples. The MLPs were then tried in combination in order to decrease the overall error rate.

One method consisted of combining the two MLPs through their output layer. The scalar product of their output vectors is measured, and the recognized class is the one for which the resulting value is the greatest. A recognition rate of 99.6 % was reached on the test set.

An alternative cooperation method is based on the fact that all the discriminant information is represented by the outputs of the hidden layer [9]. Only one output layer is then used, its inputs being taken from the outputs of the hidden layer of each of the MLPs (figure 2). So, a new training phase has to be performed, but the parameters of the hidden layers are preserved so as to extract the same discriminant information as before. As only the weights of the new output layer have to be updated, this learning phase is achieved in a restricted number of iterations. By using this second method of cooperation, a recognition rate of 99.8 % was obtained on the test set.

The difference between the results of the two combination methods is very small, so that both can be regarded as being equivalent.

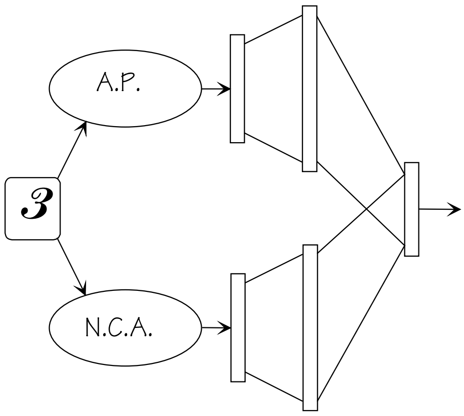

Figure 2: Cooperation through the hidden layer

Although allowing a significant reduction in error rate, combining the MLPs in parallel is also very expensive in computation time. This recognition system is 2.5 times slower than the best of the individual MLPs, i.e. the one trained on the basis of the NCA features.

The first process that was carried out in order to decrease the recognition time was to restrict the size of the MLPs that are used. As these MLPs were trained individually, their size had to increase enough to compensate for the gaps in their associated features vector. However, as has been stated, the misclassification errors produced by one of the MLPs are easily corrected thanks to the cooperation with the other one. It can then be expected that the cooperation of two MLPs of smaller size would allow a high recognition rate to be obtained. Such MLPs are already available, as a result of previous experiments, which select the most convenient number of features as well as fixing the ideal number of hidden units. So, no further learning phase is required to carry out the search for the optimal combination of MLPs.

It was possible to decrease the number of hidden units to 60 for the MLP using the AP features, and to only 30 for the one using the NCA features, while keeping an overall recognition rate of 99.8 %. For both MLPs, the number of selected features was the same as before. This leads to a recognition system which is only 40 % slower than the one using the most efficient of the MLPs alone.

6 Cascade Cooperation of Multilayer Perceptrons

A more significant improvement of the recognition time is achieved by using the notion of rejection of characters when the maximal output value of a MLP is below a pre-defined threshold. Then, it can be stated (figure 1) that the error rate on the remaining characters decreases steeply. For the MLP using the NCA features, the recognition rate is equivalent to the one obtained when both MLPs are used in parallel, for a very small rejection rate. The two MLPs can then be combined in cascade (figure 3), so that recognition is performed first by the best of the two MLPs. The parallel cooperation of the two MLPs is only used when a character is rejected by the first network.

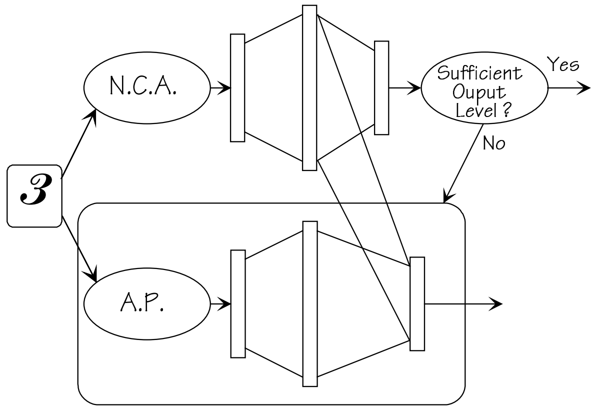

Figure 3: Cascade Cooperation of the Neural Networks

The rejection rate of the first network has to be fixed so as to limit the frequency of the use of the second network, while retaining the recognition rate at its highest value. There was no significant difference between the results of the cooperation through the hidden layer or through the output layer. The best result was obtained with a rejection rate of only 15.3 %, which allowed an overall recognition rate of 99.8 %. In other words, for a recognition rate that is equivalent to the one obtained by using both neural networks, only one of them is required, for nearly 85% of the time. Therefore, the overall recognition time is decreased by a factor 2.5, in relation to the system developed at section 5, and is also more than 40 % smaller than the one of the best of the single networks!

7 Cascade Substitution of Multilayer Perceptrons

As a cascade combination of two MLPs allows a decrease in the recognition time, a question that could be asked is whether the recognition time might be decreased further by using a cascade combination of more recognition systems. A very fast recognition system could be used initially, prior to the one developed at section 6, which would then be used only for characters rejected by the new system. Since the second recognition system is now simply a backup to the first one, there may not be any reduction in recognition time.

From the tests that were carried out, the best result was obtained by using, as the initial recognition system, a MLP containing only 10 hidden units, which used the NCA features. In order to keep a recognition rate of 99.8 %, the rejection rate had to be set to 43.5 %, which is relatively high. However, the small size of the new MLP allowed a decrease in the overall recognition time of 14 %, compared to the system developed at section 6.

A more significant improvement of the recognition time can be achieved by applying the principle of cascade cooperation to the first recognition system as well (figure 4). Such a combination of two small MLPs, each of these ones using a distinct type of features, could significantly decrease the rejection rate, at the cost of a small increase in the computation time.

The bests results were obtained by using, as first recognition system, a cascade cooperation of two MLPs containing only 10 hidden units each. So, the overall rejection rate of the first recognition system was restricted to only 8.5 %, while a recognition rate of 99.8 % was obtained. This final recognition system is 80 % faster than the system developed at section 6, and is 2.5 times faster than the best of the individual MLPs.

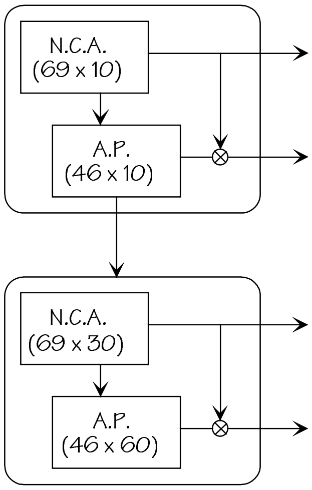

Figure 4: The final recognition system

8 Conclusion

Two kinds of features extractors for handwritten character recognition have been presented. A very efficient method for reducing the size of the features vectors has then been described. Each of the resulting features vectors has been used as input of a multilayer perceptron. Recognition rates of 97.8% and of 98.7% were reached, on handwritten digits extracted from the NIST Special Database 3. The features vectors were reduced to only 46 and 69 components respectively.

Several methods of combination of MLPs have been set out. All these methods have achieved an overall recognition rate of 99.8 % on the test set. They have then shown that cooperation of neural networks that possess distinct types of features could reduce significantly the overall error rate. Moreover, the MLPs that are combined result from the experiments that were previously performed in order to optimize the recognition process when using a single MLP. Thus, all these combination methods are very easy to carry out.

Parallel cooperation of MLPs has shown that the size of each network could be reduced without altering the overall recognition rate. Cascade cooperation of the MLPs has led to an even more significant reduction of the overall computation time. Finally, cascade substitution of two recognition systems has allowed the achievement of a very fast recognition process.

The final system consists of a combination of four MLPs, in which only the two smallest ones are used for more than 90 % of the time. This leads to a recognition process that is 2.5 faster than the one obtained when using only the best of the individual MLPs.

References:

[1] M.D. Garris and R.A. Wilkinson, « Handwritten segmented characters Database ». *Technical Report Special Database 3*, HWSC, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, February 1992.

[2] S.Impedevo, L.Ottaviano, S. Occhingro, « Optical Character Recognition - A Survey », *Int. Journal of Pattern Recognition and Artificial Intelligence*, 1991.

[3] A.d'Acierno, C.De Stefano, M. Vento « A Structural Character Recognition Method Using Neural Network » *Proc. First Int. Conf. Document Analysis and Recognition*, pp 803-811, 1991.

[4] A. Filipski and R. Flandrena, « Automated Conversion of Engineering Drawings to CAD Form », *Proceedings of the IEEE*, vol. 80, number 7, pp 1195-1209, July 1992.

[5] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, « Computer Recognition of Unconstrained Handwritten Numerals », *Proceedings of the IEEE*, vol. 80, number 7, pp 1162-1181, July 1992.

[6] H. Susuki, L.O. Chua, T. Matsumoto, « A CNN Handwritten Character Recogniser », *Int. Journal of Circuit Theory and Applications*, vol. 20, n° 5 pp 601-612, October 1992.

[7] B. Gosselin, « A Comparison of Different Feature Extractors for Handwritten Characters Recognition », *Proc. of the ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing*, Papendael, Holland, March 1994.

[8] K. Fukunaga « Introduction to Statistical Pattern Recognition », Academic Press, London, 1990.

[9] H. Bourlard & N. Morgan « Connectionist Speech Recognition - A Hybrid approach », Kluwer Academic Publishers, 1991.