Improving generalization ability of HMM/NNs based classifiers

Article · March 2015

4 authors, including:

Miguel A. Ferrer
Universidad de Las Palmas de Gran Canaria

Carlos M. Travieso
Universidad de Las Palmas de Gran Canaria

IMPROVING GENERALIZATION ABILITY OF HMM/NNs BASED CLASSIFIERS

Miguel A. Ferrer, Itziar G. Alonso*, Carlos M. Travieso*, and Aníbal R. Figueiras-Vidal†,*
* Departamento de Señales y Comunicaciones, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas de G.C., Spain. (e-mail: mferrer@dsc.ulpgc.es)
† Departamento de Tecnologías de las Comunicaciones, Universidad Carlos III de Madrid, c./ Butarque 15, 28911 Leganés-Madrid, Spain.

ABSTRACT
Standard Hidden Markov Models (HMM) have proved to be a very useful tool for temporal sequence pattern recognition, although they present a poor discriminative power. On the contrary Neural Networks (NNs) have been recognized as powerful tools for classification task, but they are less efficient to model temporal variation than HMM. In order to get the advantages of both HMMs and NNs, different hybrid structures have been proposed. In this paper we suggest a HMM/NN hybrid where the NN classify from HMM scores. As NN we have used a committee of networks. As networks of the committee we have used a Multilayer Perceptron (MLP: a global classifier) and Radial Basis Function (RBF: a local classifier) nets which drawn conceptually different interclass borders. The combining algorithm is the TopNSeg scoring method which sum the top N ranked networks normalized outputs for each class. The test of above architecture with speech recognition, handwritten numeral classification, and signature verification problems show that this architecture works significantly better than the isolated networks.

1 INTRODUCTION

The ability of hidden Markov models (HMM) of dealing with spatio-temporal sequences has been proved for several classification applications. The drawback of this competitive classification system is their reduced discriminative power, even applying alternative approaches to their classical maximum likelihood training as Maximum Mutual Information, Generalized Probabilistic Descent, etc. [1].

On the contrary, Neural Networks (NN) have been recognized as powerful machines for classification task with a good discriminative ability although are inefficient to model temporal variations even considering context, as with Time-Delay Neural Networks, or recursive NN schemes [2].

In order to get the complementary advantages of above both recognition schemes, combining HMM and NN in hybrid structures is an interesting avenue to follow: in fact, different hybrid structures have been proposed; for instance in [3], [4], etc.

In this paper, we will focus on the HMM/NN hybrid which the NN classify from HMM scores because of this arrangement looks more natural to solve spatio-temporal problems. That is to say, first the HMM deals with this spatio-temporal character time-normalizing the sequence, and, then, the high discriminative power of the NN solves the classification problem instead of the classical maxnet.

Specifically, and following previous works [5], we will introduce how the HMM+NN improves the performance of HMM classifier for several applications using different NNs. The tested NNs have been All Class One Net Multilayer Perceptron (MLP), One Class One Net MLP, Radial Basis Functions Net (RBFN), Group Method of Data Handling Net, and combination of the above mentioned global and local classifiers. The experimental results are presented in application as speech recognition, handwritten digit recognition and signature verification.

2 HMM+NN HYBRID

Usually a HMM assigns a Markov model λ_i, i={1,2 ,..,M} (where M is the number of classes) to each class c_i. In the recognition stage of the input sequence X, each HMM model λ_i estimates the a posteriori probabilities $P(X|\lambda_i)$, and the input sequence is assigned to the j class which provides the maximum score (maxnet):

$X \in \lambda_j$ *if* $j = argmax\ P(X|\lambda_i)$ i=1,2,..M

This way of doing things provides a reduced discriminative power. We stand that a way to improve the discriminative power of above scheme is to change the maxnet by a NN as in figure 1.

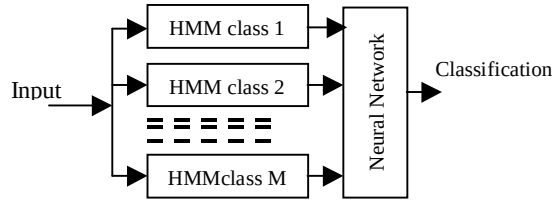

Figure 1. HMM+NN hybrid.

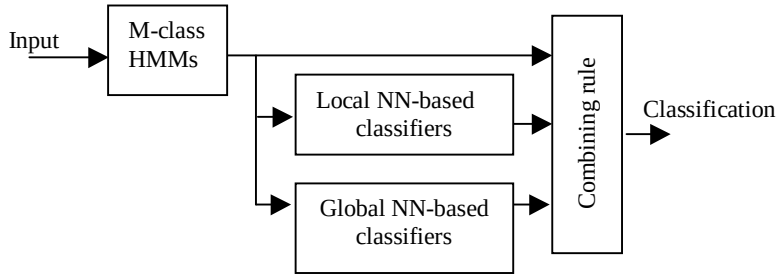

Figure 2. HMM with committee of networks

First, we suggest to use as NN a MLP (global classifier). The way of a HMM+MLP hybrid recognizer can be constructed, as suggested in [4][5], is applying the a posteriori probabilities $P(X|\lambda_i)$ as inputs of an MLP; assuming we select an N units hidden layer architecture, the output is calculated as

$$O_1(i|X) = f\left\{\sum_{k=1}^{N} w_2(i,k) f\left[\sum_{j=1}^{M} w_1(k,l) P(X|\lambda_j) + \theta_1(k,j)\right] + \theta_2(i,k)\right\}$$

where $w_2(i,k)$ is the weight from the kth hidden neuron to the ith output neuron, $w_1(k,j,l)$ is the weight from the lth state of the jth HMM model to the kth hidden neuron, $\theta_1(k,j,l)$ and $\theta_2(i,k)$ are the constant biases of the hidden and output layer neurons, and f is the global activation function, usually a sigmoid.

Second, in order to test the HMM+NN structure with a local classifier, we have used as NN a Gaussian Radial Basis Function Network (RBFN); its output is given by

$$O_2(i|X) = \sum_{k=1}^{N} w(i,k) \exp\left\{-\sum_{j=1}^{M} \frac{(P(X|\lambda_j) - \mu(k,j))^2}{2\sigma^2(k,j)}\right\}$$

being $w(i,k)$ the weights from the kth Gaussian neuron to the ith output, and $\mu(k,j,l)$ and $\sigma^2(k,j,l)$ are the jth, lth components of the kth Gaussian neuron centroid and spread ("mean" and "variance"), respectively.

3 HMM+COMMITTE OF NETWORKS

A drawback of the above HMM+NN hybrid it is the common practice of training many different candidate networks and then to select the best on the basis of performance on an independent validation set, and to keep only this network and to discard the rest. There are two disadvantages with such an approach: First, all of the involved effort in training the remaining networks is wasted. Second, the generalization performance on the validation set has a random component due to the noise on the data, and so the network which has best performance on the validation set might not be the one with the best performance on new test data.

These drawbacks can be overcome by combining the postprocessing networks of the HMM to form a committee [6]. The importance of such an approach is that it can lead to significant improvements in the classification, while involving little additional computational effort. That improvement depends on the care taken in two tasks: first, the choice of the networks that compound the committee (decorrelated networks), and second, the combining algorithm of the networks.

A way of reducing the correlation among networks in a committe is combining networks that drawn conceptually different interclass borders. So, if we adequately combine the outputs of HMM with outputs of both above mentioned local and global classifiers, we expect significantly improved results and we could say that we are working with a kind of global-local classifier (see figure 2).

In order to combine the networks' output and obtain the scores from which decide the input class, we normalize the output of all the networks among [0 1] and apply a combining algorithm. Several combining algorithms have been tested.

The *NoSeg* score of each class is defined as the best score over all networks. The *SumSeg* score is the sum over all networks of each class's normalized score. The *TopNSeg* scoring method is to sum the top N ranked networks normalized outputs for each class. Note that the networks selected in the sum vary by class. It seems reasonable that contaminating networks my have lower scores for the true class, in which case TopNSeg eliminates contaminating scores. *Top1Seg*, a special case of TopNSeg is the maximum network scores for each class. A variant on the TopNSeg score, *TopSeg-MtoN*, is defined as the sum of the ranked network scores M to N. Unwanted high scoring networks can happen when an anomalous event scores significantly better on one class's networks than other class's networks even though the unnormalized scores of all models are relatively low [7].

After test all of them, the best results were obtained with TopNSeg scoring method and we have used this combining method.

4 INSUFFICIENT TRAINING DATA

Experimentally, we realized that when the training sequence is insufficient, the NN postprocessor does not improve the HMM classification ratio. It may be because the HMM

classification ratio over the training sequence is greater than 99% and when training the NN postprocessor it does not receive information about HMM mistakes. If we want to obtain a convenient generalization, it is necessary to allow that the NN perceives the potential sources of misclassification present at the HMM.

In order to alleviate this drawback, we used a new training method based on an idea similar to that under the classical cross-validation training methods, as well as very easy to apply: to train the HMM with a subset of the available training sequence, and, after this, to train the NN with the whole training sequence. Since this method allows the NN to deal with inputs not used to design the HMM, it is reasonable to assume that the NN receives more information about potential HMM mistakes: thus, increasing its capability of correcting them. This is the reason according to which we call this combined training "correcting mistakes" method.

5 EXPERIMENTAL RESULTS

The paper will present experimental results with three application: isolated word recognition, handwritten digit recognition and, for insufficient training data case, signature verification.

For speech recognition, the database used in the experiments consisted of 11 repetitions of the Spanish digits spoken by 133 speakers (chosen according their sex and age for a fair representation). Each digit was modeled by a Discrete HMM. The vector quantizer has 64 codewords. Each HMM is a left-to-right model of 10 states and has been trained and tested using the Baum-Welch and Viterbi algorithm. The NN has 75 hidden neurons and has been trained according to the resilient backpropagation algorithm and cross-validation with active pattern selection. The RBF has 256 gaussian radial basis functions calculated by modified k-means algorithm [5].

For unconstrained handwritten numerals, the database used in the experiments consisted of 12 repetitions of each numeral written by 90 writers; the writers were chosen according to their sex, age, and handwriting for a fair representation. The numerals turned to one-dimensional vector by following the contour line of the image [8]. Each numeral was modeled by a HMM. Each HMM is a left-to-right model of 50 states and have been trained and tested using the Baum-Welch and Viterbi algorithm. Trained in a similar way that in above case, the NN has 80 hidden neurons and the RBF with 256 neurons [9].

For signature classification, the database used in the experiments consisted of 24 repetitions of the signatures of 60 writers; so we have just 24 repetitions of each class. The digital processing of the signatures is similar to that applied to handwritten digits. In this case each HMM is a left-to-right model of 35 states.[14]. NN has not been used with NN because the ratio NN_weigths divided by length_of_training_sequence is too high. The used RBF has 100 Gaussian radial basis functions [10].

In order to reliable test of above schemes, the results have been calculated as follow: for each application, we have trained the different systems with a subset of the entire database chosen randomly (30% for speech recognition, 40% for numeral recognition, and 60% for signature classification) and tested the system with the other disjoint set. This procedure is repeated 10 times with different randomly selected training and test sets and the results are averaged. For corrective training, the HMM is trained with the 50% of training sequence. The results obtained can be seen at tables I, II and III.

From below tables, we can see that isolated NNs not always improve the performance of the HMM classifier. The real improvement is obtained when combining HMM+ MLPs+ RBFs's outputs. This improvement use to be greater with corrective mistake than with standard training method. An additional advantage of corrective training method is that the variance of the scores obtained over the ten different experiments is lesser than the variance obtained with standard training method.

The above along with the fact that the computational load of corrective training method is lesser than the computation need of standard training method make the HMM + committee of networks architecture with corrective training a serious alternative to HMM with standard training based recognizers.

6 CONCLUSIONS

In this paper, we have shown that there is room to obtain significant improvements in recognition performance using HMM+NN hybrids, without the need of an overall training, if the idea of "generalization" is kept in mind when training both elements: i.e., if it is allowed that the NN sees enough "mistakes" as to have the possibility of learning how to correct them. Even a simple procedure, based on adding samples to the training set for the HMM when training the NN, provides support to this approach, as it can be seen in the results of two simple experiments we have carried out.

TABLE I. AVERAGE RECOGNITION RATES FOR THE HMM, HMM+NN AND HMM+COMMITTEE OF NETWORKS FOR ISOLATED SPEECH RECOGNITION APPLICATION.

Training method	HMM	HMM+MLP	HMM+RBF	Combining
Standard training	96.0±0.7	96.3±0.8	95.7±0.7	96.7±0.6
Corrective mistake	96.3±0.5	96.1±0.5	95.7±0.5	96.9±0.4

The results are given as average ± variance

TABLE II. AVERAGE RECOGNITION RATES FOR THE HMM, HMM+NN AND HMM+COMMITTEE OF NETWORKS FOR HANDWRITTEN DIGIT CLASSIFICATION APPLICATION.

Training method	HMM	HMM+MLP	HMM+RBF	Combining
Standard training	95.1±0.8	96.1±0.5	94.9±0.7	95.8±0.6
Corrective mistake	93.8±0.8	95.0±0.8	94.0±0.6	95.9±0.6

The results are given as average ± variance

TABLE III. AVERAGE RECOGNITION RATES FOR THE HMM, HMM+NN AND HMM+COMMITTEE OF NETWORKS FOR SIGNATURE VERIFICATION

Training method	HMM	HMM+RBF	Combining
Standard	84.4±3.5	80.8±9.0	82.4±2.8
Corrective mistake	82.2±2.3	84.6±3.4	85.2±2.0

The results are given as average ± variance

This conception of training hybrid recognizers open multiple research questions: not only about selecting the adequate NN architectures, sizes, and objectives for each particular recognition system, but also on what are the most convenient input values (taken from the previously designed HMM) for these NN, and what kind of training must be used to get a high generalization; even it can be convenient to rethink on how to train the HMM, considering that its variables will not be applied to decide, but to serve as data for a classifier.

REFERENCES

[1]. H.Boulard, H.Hermansky, N.Morgan, "Towards Increasing Speech Recognition Error Rates", *Speech Communications*, vol.18, no.3. pp.205-231, May 1996.

[2]. N. Morgan and H. Borlard, "neural Networks for Statistical recognition of Continuous Speech", *Proceedings of the IEEE*, vol. 83, no. 5, pp. 742-770, May 1995.

[3]. A.Kundu, G.C.Chen, "An Integrated Hybrid Network and Hidden Markov Model Classifier for Sonar Signals", *IEEE Trans. On Signal processing*, vol. 45, no.10, pp. 2566-2570, October 1997.

[4]. S.B.Cho, "Neural-Network Classifier for Recognizing Totally Unconstrained Handwritten Numerals", *IEEE trans. on Neural Networks*, vol.8, no.1, pp.43-53, January 1997.

[5]. Itziar G. Alonso, Carlos M. Travieso, Miguel A. Ferrer, Aníbal R. Figueiras, "A Comparative Study of Different Hybrid HMM/NN Classifiers for Speech Recognition", *Proceedings of the IASTED International Conference, Signal and Image Processing*, Las Vegas, Nevada-USA, paper nº 281-125, October 1998.

[6]. Amanda L.C.Sharkey, "On Combining Artificial Neural Nets", *Connection Science*, vol. 8, Number 3&4, pp.299-314, December 1996.

[7]. H.Gish and M.Schmidt, "Text-Independent Speaker identification", in *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18-32, October 1994.

[8]. Ciro R. Morales, Carlos M Travieso, Itziar G. Alonso, Miguel A. Ferrer, "Algoritmo Secuencial de Esqueletización para un Sistema de Reconocimiento de Caracteres", *XIII Simposium Nacional de la Unión Científica Internacional de Radio*, páginas 199-200, Santiago de Compostela, Septiembre 1999.

[9]. Carlos M Travieso, Ciro R. Morales, Itziar Goreti Alonso, Miguel Ángel Ferrer, "Handwritten Digits Parameterisation for HMM Based Recognition", *VII IEE Int. Conf on Image Processing and its applications*, páginas 770 a 774, Manchester, UK, Julio de 1999.

[10]. Jose L. Camino, Carlos M. Travieso, Ciro R. Morales, Miguel A. Ferrer, "Signature Classification by Hidden Markov Models", *33rd Annual IEEE International Carnahan Conference on Security Technology*, pp. 481-585, Madrid, España, Octubre de 1999.