ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Non-linear gating network for the large scale classification model CombNET-II

Mauricio Kugler, Toshiyuki Miyatani
Susumu Kuroyanagi, Anto Satriyo Nugroho and Akira Iwata *

Department of Computer Science & Engineering
Nagoya Institute of Technology, Gokiso-cho, Nagoya, 466-8555 - Japan

Abstract. The linear gating classifier (stem network) of the large scale model CombNET-II has been always the limiting factor which restricts the number of the expert classifiers (branch networks). The linear boundaries between its clusters cause a rapid decrease in the performance with increasing number of clusters and, consequently, impair the overall performance. This work proposes the use of a non-linear classifier to learn the complex boundaries between the clusters, which increases the gating performance while keeping the balanced split of samples produced by the original sequential clustering algorithm. The experiments have shown that, for some problems, the proposed model outperforms the monolithic classifier.

1 Introduction

The large scale classification model CombNET-II, proposed by Hotta *et al.* [1], is a divide-and-conquer based method able to deal with databases of thousands of categories. It has presented several good results in Chinese character (Kanji) recognition and some other specific applications. In its basic form, the CombNET-II is composed by a gating network (stem network) and many expert networks (branch networks). The stem network is a modified Vector Quantization (VQ) based sequential clustering called Self Growing Algorithm (SGA), while the branch networks are basically independent Multilayer Perceptrons (MLP). Essentially, the stem network is used to divide the feature space in R Voronoi subspaces, each in turns becomes the training data for the MLPs.

For large scale problems, however, the use of raw data on the stem network training causes the classes to be shattered among the clusters. This creates very unbalanced problems for the branch networks, and each cluster will end up containing a large number of classes. These two problems can cause complex and slow branch network training. A solution for this is the use of the average of each class samples on the SGA algorithm training. This procedure, apart from reducing the stem network training time, also avoids the classes to be split among the clusters. This reduces the number of classes per cluster and improves the balance of samples of different classes inside each branch network.

However, the averaged data does not represent thoroughly the real data, specially for complex distributions. If the real training samples were applied

*The first author is supported by the Ministry of Education, Culture, Sports, Science and Technology, Government of Japan, and also by a grant from the Hori Information Science Promotion Foundation, Japan.

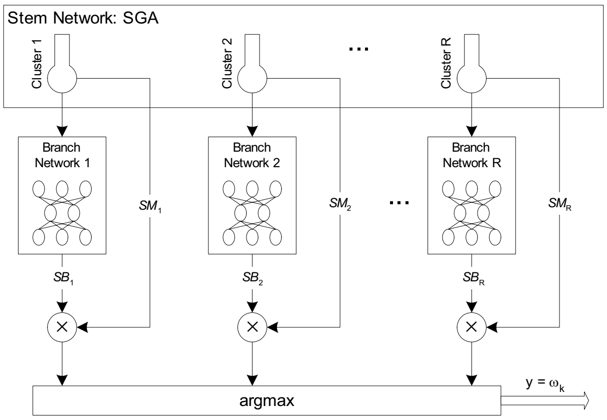

Fig. 1: CombNET-II structure

to the stem network trained with the averaged samples, a bad performance could be expected. This problem tends to deteriorate with increasing number of clusters, because the feature space learned by each branch network starts to differ more and more from the feature space represented by the corresponding stem cluster. Clearly, there is a compromise between the stem and the branch networks performance.

This paper proposes a new solution that eliminates this compromise, which increases the stem network performance while keeping the advantages of the use of averaged data. An independent MLP is used to represent the complex boundaries between the clusters generated by the use of averaged data, which increases the stem network performance without interfering on the balance of the branch networks feature space.

Non-linear algorithms had already been used as gating networks for large scale models. Collobert, Bengio and Bengio [2] used a MLP gating in their large scale model. However, in their approach, the training data splitting starts randomly, and is iteratively redefined based on the expert networks performance. This requires the gating to be retrained on each iteration, hence making the procedure very time consuming. Waizumi *et al.* [3] presented a new rough classification network for large scale models based on a hierarchy of Learning Vector Quantization (LVQ) neural networks. However, no definite result from the application of their gating network in a complete large scale model was presented. The method proposed in this paper uses a hybrid gating, in which the non-linear algorithm learns the data splitting generated by the unsupervised clustering algorithm.

2 Large Scale Classifier CombNET-II

As explained before, the CombNET-II is composed by a gating network, called the “stem” network, and many expert networks, called the “branch” networks, with its basic structure shown in Figure 1.

The stem network is a VQ based sequential clustering with some modifications to control the balance between the clusters, which replaces the Self Organizing Map (SOM) used in the original CombNET [4]. The branch networks are independent MLP networks, and have their classification results weighted by the stem network scores as described in equation (1).

$$y = \omega_k \left| S_k = \arg\max_j \left(SM_j^{\gamma} \cdot SB_j^{1-\gamma} \right) \right. \quad (1)$$

where SM_j is the similarity (usually, the normalized dot product) between the unknown sample $\mathbf{x}$ and the j^{th} cluster $\mathbf{m}_j$, SB_j is the maximal score among the output neurons of the j^{th} branch network and ω_k is the k^{th} possible category, $k = 1, \ldots, K$. The exponent γ is a weighting parameter ($0 \leq \gamma \leq 1$) that dictates which network (stem or branch) plays the major role in the classification.

In its normal form, the SGA has no control about the number of classes in each cluster or the balance of the classes inside each cluster. With the use of averaged data, this control is not necessary (considering that the original classes are nearly balanced), as the control of the clusters size already regulates the number of classes on it. The stem network classification accuracy, however, tends to deteriorate. The next section presents the proposed strategy for improving the stem network performance trained with averaged data.

3 Proposed Model

Instead of changing the clustering result in order to search for different data splits that could improve the stem network result without sacrificing the branch networks performance, this paper proposes the use of a non-linear algorithm to learn the complex boundaries between the clusters generated by the use of averaged data on the SGA training.

At first, the SGA algorithm is trained using the averaged samples $\bar{\mathbf{x}}_k$ of each k^{th} class. The raw data is then split using the obtained cluster belonging information: $\mathbf{x}_i \in \mathbf{m}_j \leftrightarrow [y_i = k,\ \bar{\mathbf{x}}_k \in \mathbf{m}_j]$. The samples belonging to cluster $\mathbf{m}_j$ are used to train the j^{th} branch network. Independently, the raw data is also relabeled by the clustering information: $y'_i = j \leftrightarrow [y_i = k,\ \bar{\mathbf{x}}_k \in \mathbf{m}_j]$, where y_i and y'_i are respectively the original and the new label of the i^{th} sample. The relabeled raw data is the training data for the non-linear gating network.

The relabeled data characteristics suggest the use of MLP as the stem non-linear algorithm, although any other method could have been used. To avoid misunderstandings between the branch MLPs and the stem network MLP, the latter will be abbreviated as S-MLP. The S-MLP training is independent of the branch MLP networks and can be made in parallel.

On the recognition stage, the clustering result of the SGA is no longer needed. The unknown sample is inputted directly on the S-MLP and, instead of the linear stem network similarity, each SM_j will correspond to the S-MLP j^{th} output neuron result, to be multiplied by the correspondent SB_j value in equation (1). The final structure of the proposed model is shown diagrammatically in Figure 2.

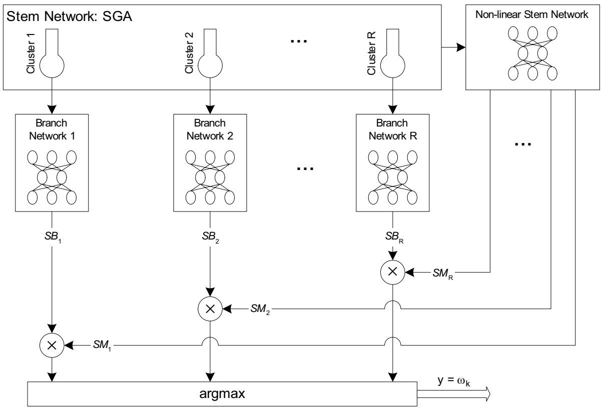

Fig. 2: Proposed model structure

4 Experiments

The experiments showed in this paper intend to verify the proposed model's performance gain for the cases where the SGA is trained with averaged data. That is the case of large scale problems, for which it is impracticable to train the stem network with raw data. Therefore, even though the databases used in this paper experiments can not be considered large (and so allowing the stem network to be trained with raw data), they can properly represent the problem of using averaged data in the SGA. The medium size experiment allows extensive optimization of the parameters, which provides a more thorough understanding of the models' behaviors.

The same linear stem network and branch networks were used for both models. The MLP neural networks (both branch MLP and S-MLP) were trained until the error was smaller than 10^{-4} or the iteration number exceeds 10^3, with learning rate equals to 0.9, momentum 0.1 and sigmoidal activation function slope 0.1, while the number of hidden neurons and the γ parameter were optimized for each experiment realization.

Two databases were used to verify the performance of the proposed model: *Alphabet* and *Isolet*. The *Alphabet* database consists of the roman alphabet characters subset of the JEITA-HP database [1] dataset A. The first 200 samples of each character from A to Z were selected for the experiment, with 150 for training (3900 samples) and 50 for testing (1300 samples), preprocessed by a 256 dimensions Local Line Direction (LLD) feature extraction method [5]. The *Isolet* database, obtained from the UCI repository [6], contains 26 categories representing spoken names (in English) of each letter of the alphabet. Each letter was spoken twice by each of the 30 speakers, totalizing 7800 samples (3 of them are missing), divided in 6238 samples for training and 1559 for testing, with 617 features per sample. For both databases, the stem network was trained with several parameters in order to obtain increasing number of clusters, with

[1] Available under request from
http://tsc.jeita.or.jp/TSC/COMMS/4_IT/Recog/database/jeitahp/index.html

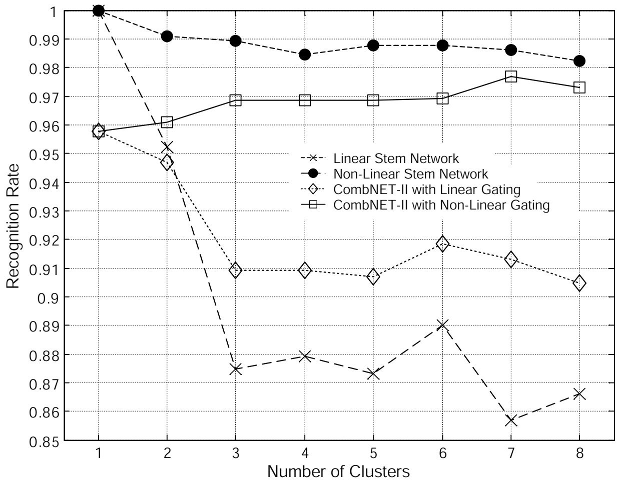

Fig. 3: Results for the *Alphabet* database

the best possible balance of number of classes between them and avowing single class clusters.

Figures 3 and 4 depicts the classification accuracy for both databases, in which the crosses' dashed line represents the linear stem network, the dark circles' dashed line represents the non-linear stem network, the diamonds' dotted line represents the standard CombNET-II overall results (by applying equation (1) to combine stem and branches) and the squares' solid line represents the overall results of CombNET-II with the non-linear gating. The gating classification accuracy is calculated verifying if the winning branch network is the same that contains the unknown sample's class training samples.

There was a significant improvement in the error rate by the use of the non-linear gating, especially for high number of clusters. For the *Alphabet* database, even the monolithic classifier was outperformed for all cases. For the *Isolet* database, there was also a significant improvement.

5 Discussion and Conclusions

The results shown in Figures 3 and 4 confirm the superiority of the proposed method. As expected, a considerably higher performance of the stem network was obtained by the use of a non-linear classification algorithm. The stem network error rate reduction lies between 80.6% and 91.4% for the *Alphabet* database and between 63.9% and 97.3% for the *Isolet* database, in comparison with the linear stem network. Consequently, the CombNET-II error rate was reduced by up to 73.4% and 40% for the *Alphabet* and *Isolet* databases respectively.

The independence of the non-linear stem network, due to the use of the same clustering information used to split the data for the branch networks, makes the proposed model very flexible and easy to implement and train. However, for

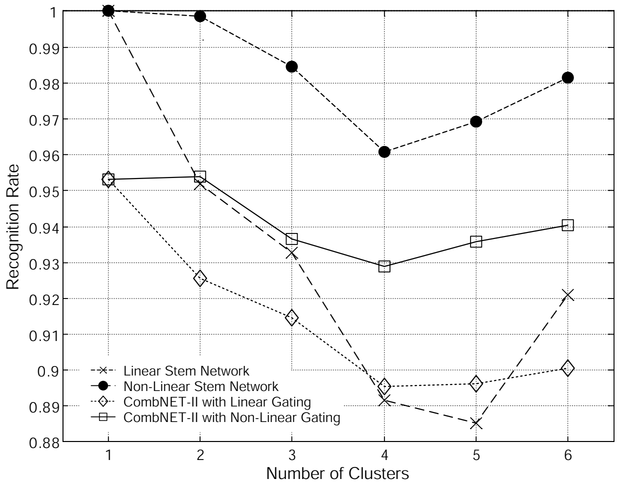

Fig. 4: Results for the *Isolet* database

very large databases, the training time for the stem network can be a bottleneck of the system, as it uses the whole raw training data (even it being relabeled for a small number of categories).

Future works include methods for reducing the training time of the non-linear stem network through reduction in the size of its training data, in order to apply the proposed model in larger databases. Also, the use of other kinds of (dis)similarity measurements on the stem network and other types of non-linear algorithms than the S-MLP will also be evaluated for performance improvement.

References

[1] Kenichi Hotta, Akira Iwata, Hiroshi Matsuo, and Nobuo Susumura. Large scale neural network CombNET-II. *IEICE Transactions on Information & Systems*, J75-D-II(3):545–553, March 1992.

[2] Ronan Collobert, Samy Bengio, and Yoshua Bengio. Scaling large learning problems with hard parallel mixtures. *International Journal on Pattern Recognition and Artificial Intelligence*, 17(3):349–365, 2003.

[3] Yuji Waizumi, Nei Kato, Kazuki Saruta, and Yoshiaki Nemoto. High speed and high accuracy rough classification for handwritten characters using hierarchical learning vector quantization. *IEICE Transactions on Information & Systems*, E83-D(6):1282–1290, June 2000.

[4] Akira Iwata, Takashi Touma, Hiroshi Matsuo, and Nobuo Suzumura. Large scale 4 layered neural network “CombNET”. *IEICE Transactions on Information & Systems*, J73-D-II(8):1261–1267, August 1990.

[5] H. Kawajiri, T. Yoshikawa, J. Tanaka, A.S.Nugroho, and A. Iwata. Handwritten numeric character recognition for facsimile auto-dialing by large scale neural network CombNET-II. In *Proceedings of the 4th International Conference on Engineering Application of Neural Networks*, pages 40–46, Gibraltar, June 1998.

[6] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.