PAPER

CombNET-III with nonlinear gating network and its application in large-scale classification problems

Mauricio KUGLER[†a)], *Nonmember*, Susumu KUROYANAGI[†b)],
Anto Satriyo NUGROHO[††c)], *and* Akira IWATA[†d)], *Members*

SUMMARY Modern applications of pattern recognition generate very large amounts of data, which require large computational effort to process. However, the majority of the methods intended for large-scale problems aim to merely adapt standard classification methods without considering if those algorithms are appropriated for large-scale problems. CombNET-II was one of the first methods specifically proposed for such kind of a task. Recently, an extension of this model, named CombNET-III, was proposed. The main modifications over the previous model was the substitution of the expert networks by Support Vectors Machines (SVM) and the development of a general probabilistic framework. Although the previous model's performance and flexibility were improved, the low accuracy of the gating network was still compromising CombNET-III's classification results. In addition, due to the use of SVM based experts, the computational complexity is higher than CombNET-II. This paper proposes a new two-layered gating network structure that reduces the compromise between number of clusters and accuracy, increasing the model's performance with only a small complexity increase. This high-accuracy gating network also enables the removal the low confidence expert networks from the decoding procedure. This, in addition to a new faster strategy for calculating multiclass SVM outputs significantly reduced the computational complexity. Experimental results of problems with large number of categories show that the proposed model outperforms the original CombNET-III, while presenting a computational complexity more than one order of magnitude smaller. Moreover, when applied to a database with a large number of samples, it outperformed all compared methods, confirming the proposed model's flexibility.

key words: *large-scale classification problems, support vector machines, gating networks, divide-and-conquer*

1. Introduction

Modern applications of pattern recognition generate very large amounts of data, which require large computational effort to be processed. Human-computer interface (speech, handwriting and gesture recognition), bioinformatics, object recognition and spam mail detection are a few examples of such kind of application. Due to these applications, large-scale classification methods have being receiving an increasing attention.

However, the majority of the methods intended for large-scale problems aim to directly adapt standard classification methods, for example, reducing training time by decomposition strategies. These methods do not consider if the standard classifier structure is actually appropriated for large-scale problems.

Classifiers ensembles are often used in cases where a single classifier cannot properly represent the solution [1]. Mixture of experts [2] is an ensemble technique in which different classifiers specialize in different regions of the input space. When the data is split off among the classifiers, the resulting structure is usually referred as a divide-and-conquer classification model.

Several methods based on this principle had been proposed. Methods using Multilayer Perceptron (MLP) based experts are described in [3]–[7], mainly dedicated to problems with large number of categories. Those methods, however, implement several heuristics in order to reduce computational complexities related to training and classification, which complicate their use as components of other systems. Support Vector Machines (SVM) based divide-and-conquer classifiers, described in [8]–[10], were applied on binary problems with large number of samples.

The model proposed on this paper is based on the CombNET model, first introduced in [11]. In order to solve problems of unbalance among the experts, the gating network was modified and the model extended to CombNET-II in [12]. CombNET-II presents a simpler and more flexible structure than the models in [3]–[7]. The latest extension, CombNET-III [13], substituted the original MLP based experts by multiclass SVM. It also implements a probabilistic framework, enabling its direct application as part of other systems. A more detailed description of CombNET-III will be made in section 2.

The main objectives of this paper are the improvement of CombNET-III performance by the use of a nonlinear gating network and the reduction of its classification computational complexity, pointed in [13] as a main concern for further developments. The use of a nonlinear gating network with higher accuracy permits the elimination of less confident experts on the

Manuscript received January 1, 200?.
Manuscript revised January 1, 200?.
Final manuscript received January 1, 200?.

[†]The authors are with the Department of Computer Science & Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan.

[††]The author is with Agency for the Assessment and Application of Technology (PTIK-BPPT), Jalan MH.Thamrin No.8 Jakarta, 2nd Bld. 21F, Jakarta-Indonesia 10340.

a) E-mail: mauricio@kugler.com
b) E-mail: bw@nitech.ac.jp
c) E-mail: asnugroho@inn.bppt.go.id
d) E-mail: iwata@nitech.ac.jp

decoding phase, reducing considerably the number of required calculations. Moreover, this paper introduces a new strategy for reducing the number of Kernel function evaluations performed by the multiclass SVM experts when evaluating an unknown sample. This strategy can also be applied in stand-alone multiclass SVM implementations.

Furthermore, the increased accuracy of the gating network enables the application of the proposed model in problems with large number of samples, which were not a concern in past CombNET model development works. This paper aims to evaluate the proposed model performance when applied to these problems, in comparison to other recently proposed large-scale methods.

The organization of the paper goes as follows: a more detailed revision of CombNET-III is presented in Section 2, and Section 3 introduces the proposed model, its modifications and new characteristics. Section 4 presents experiments with the new model and some comparisons with previous results, and Section 5 concludes the paper with analysis of the results and suggests possible future extensions.

2. Large-Scale Classifier CombNET-III

The CombNET-II model proposed by Hotta *et al.* [12] is a large-scale classifier that follows the classic structure of divide-and-conquer methods: a gating network and many experts classifiers, called respectively "stem" network and "branch" networks in the original references. The branch networks are MLP trained by gradient descent, while the stem network is a modified VQ based sequential clustering algorithm, called Self Growing Algorithm (SGA), developed to solve the problem of unbalanced clusters generated by the Self-Organizing Map (SOM) used in the original CombNET [11]. The basic SGA algorithm is described in Figure 1, in which ℓ is the number of samples, R is the current number of clusters, $\mathbf{x}_i$ is the i^{th} sample, ν_j is the j^{th} cluster reference vector, Θ_s is the similarity threshold, Θ_P is the inner potential threshold, h_j is the j^{th} cluster inner potential and $sim\left(\nu_j, \mathbf{x}_i\right)$ represents the similarity measurement between the i^{th} sample and the j^{th} cluster.

```
Make $\nu_1 = \mathbf{x}_1$, $h_1 = 1$ and $R = 1$
for $i \in \{2 \ldots \ell\}$
    Find $\nu_c$ so that:
        $sim\left(\nu_c, \mathbf{x}_i\right) = \max_j \left[sim\left(\nu_j, \mathbf{x}_i\right)\right]$
    if $sim\left(\nu_c, \mathbf{x}_i\right) < \Theta_\mathrm{s}$
        $R = R + 1$, $\nu_R = \mathbf{x}_i$, $h_R = 1$
    else
        $\nu_c^{new} = \nu_c^{old} - h_c^{-1}\left(\mathbf{x}_i - \nu_c^{old}\right)$
        $h_c = h_c + 1$
        if $h_c > \Theta_\mathrm{P}$
            Divide $\nu_c$ in $\nu_c'$ and $\nu_{R+1}$ so that:
                $\left|h_c - h_{R+1}\right| \leq 1$
        end if
    end if
end for
do Update the clusters (with necessary divisions)
until No significant changes in any clusters
```

Fig. 1 Self Growing Algorithm (SGA)

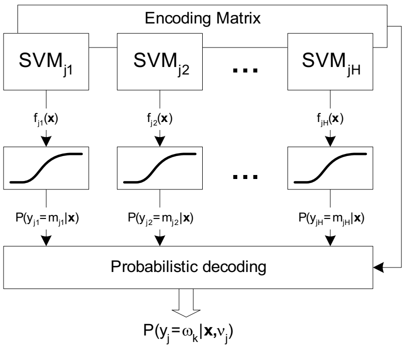

Fig. 2 SVM branch network structure

Kugler *et al.* [13] recently presented an extension of this model, called CombNET-III, which substitutes the MLP branch networks by multiclass Support Vector Machines (SVM) based branch networks and introduces a new probabilistic framework for combining the branch networks' outputs. The SVMs uncalibrated outputs were moderated by a sigmoid function in order to generate class posterior probabilities, using Platt's approach [14]. The branch network structure is shown in Figure 2.

The probabilistic framework for calculating the branch networks outputs and the CombNET-III final posterior probability are defined in [13] as follows. Given an unknown sample $\mathbf{x}$, the j^{th} branch network posterior probability of class ω_k is given by:

$$P\left(\omega_k \left|\mathbf{x}, \nu_j\right.\right) = \frac{\sum\limits_{h:m_{k,h}\neq 0} P\left(y_{k,h} = m_{k,h}\left|\mathbf{x}\right.\right)}{\sum\limits_{h=1}^{H}\left|m_{k,h}\right|} \tag{1}$$

where $\mathbf{M}^{K\times H}$ is the coding matrix, with $m_{k,h} = \{-1, 0, +1\}$, K is the number of classes, H is the number of classifiers, and ν_j is the j^{th} cluster. The final posterior probability of the class ω_k given an unknown sample $\mathbf{x}$ is given by:

$$P\left(\omega_k\left|\mathbf{x}\right.\right) = c\prod_{j=1}^{R}\left[P\left(\nu_j\left|\mathbf{x}\right.\right)^{\gamma}P\left(\omega_k\left|\mathbf{x}, \nu_j\right.\right)^{1-\gamma} + \right. \\ \left. + \frac{1 - P\left(\nu_j\left|\mathbf{x}\right.\right)^{\gamma}}{2}\right] \tag{2}$$

where the term c before the product is used to adjust the probabilities scale in order to ensure they are calibrated, summing to unity, R is the total number of clusters, ν_j is the j^{th} cluster and γ is a weighting factor between the cluster posterior probability and the

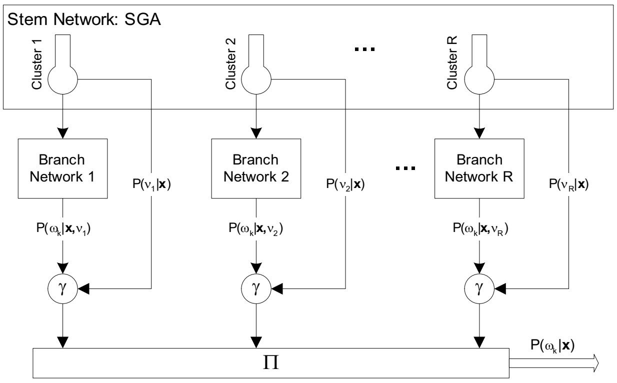

Fig. 3 CombNET-III structure

branch networks class probabilities. The final structure of the CombNET-III is shown diagrammatically in Figure 3.

CombNET-III outperformed both CombNET-II and a single multiclass SVM, while presenting much smaller computational complexity than the last. It also presented a much smaller training time. Notwithstanding these advantages, the CombNET-III model still presents some limitations similarly to CombNET-II. Although presenting less interference between the branch networks (as the SVM outputs with gaussian kernel tends to zero for outlier samples), the performance of CombNET-III still depends on the gating network accuracy, which decrease rapidly with increasing number of clusters. Consequently, large number of clusters, required by very large-scale problems, still cannot be used. Another disadvantage of CombNET-III is its high classification computational complexity. The use of SVM as the expert classifiers considerably increased the number of required calculations when comparing to CombNET-II. Even this complexity is much smaller than the single multiclass SVM, it can be a limiting factor for the application of CombNET-III in real world problems. These are the two problems addressed on this paper by the proposed model introduced in the next section.

3. Proposed Model

3.1 Nonlinear gating network

The standard SGA algorithm is an unsupervised procedure. Thus, it does not consider the label of the samples when clustering the data and has no controlling mechanism for the number of categories neither their balance inside each cluster. Therefore, for large-scale problems with large number of categories, the use of raw data on the stem network training causes the samples belonging to categories with complex distributions to be shattered among the clusters. This creates very unbalanced problems for the branch networks, and some cluster can end up containing a large number of classes. These two problems can make the branch networks training complex and slow, reducing also the overall classifier performance.

In order to solve this problem, previous works proposed the use of the average of each class samples on the SGA algorithm training, instead of the raw data samples [12], [15]. This straightforward procedure, apart from reducing the stem network training time, also avoids the classes to be split among the clusters. Hence, it 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. In the case of complex distributions, several samples that belongs to a certain cluster can present a higher similarity with some other neighbor clusters. This problem tends to deteriorate with increasing number of clusters, because the samples subspace learned by each branch network starts to differ more and more from the subspace 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, increasing the stem network performance while keeping the advantages of the use of averaged data.

The main reason for the standard stem network of CombNET-II and CombNET-III to present poor performance with increasing number of clusters is the use of a single reference vector for representing each of the clusters. These reference vectors can only define linear hyperplanes between the clusters, thus being unable to represent the true complex boundaries generated by the use of averaged data. Even though several methods which implement simple VQ based gating networks do not present any mechanism for controlling the balance among the clusters [3]–[5], [16], they present a higher gating network accuracy due to the use of multiple reference vectors to represent the clusters. The use of multiple reference vectors, although increasing the gating computational complexity, defines complex nonlinear boundaries between the clusters, which are a more faithful representation of the samples subspaces learned by the branch networks.

In order to obtain a high accuracy gating network while keeping the clusters' balance control, the proposed method generates multiple reference vectors for each of the clusters, according to the algorithm shown in Figure 4, in which the SGA algorithm from Figure 1 is used as a subroutine, ℓ_k is the number of samples belonging to the k^{th} category, $\varsigma_{j,s}$ is the s^{th} reference vector of the cluster ν_j, S_j is the number of reference vectors representing the j^{th} cluster and the notation $\bar{\mathbf{x}}_k \in \nu_j$ is defined as:

$$\mathbf{x}_k \in \nu_j \leftrightarrow sim\left(\mathbf{x}_k, \nu_j\right) > sim\left(\mathbf{x}_k, \nu_c\right) \forall c \neq j \qquad (3)$$

According to the Figure 4, after the SGA is used on the averaged data (similarly to the original CombNET-III), each j^{th} cluster's correspondent raw data is independently clustered, again using the SGA, generating a set of reference vectors $\varsigma_{j,s}$, where $s = 1 \ldots S_j$. The cluster posterior probability for an unknown sample $\mathbf{x}$ then becomes:

$$P\left(\nu_j \left|\mathbf{x}\right.\right) = \max_{\varsigma_{j,s} \in \nu_j} P\left(\varsigma_{j,s} \left|\mathbf{x}\right.\right) \tag{4}$$

and can be directly applied in equation (2). On the decoding phase, the reference vectors $\{\nu_1 \ldots \nu_R\}$ generated from the averaged data are no longer used. The proposed model is diagrammatically illustrated in Figure 5. From this point, the original self growing algorithm and the new two-layered structure will be referenced respectively SGA-I and SGA-II.

Nonlinear algorithms had already been used as gating networks for large-scale models. These strategies, as well as the proposed SGA-II, are composed by two stages: some clustering strategy divides the data and a non-linear classifier learns the generated hyperplanes. For instance, Collobert, Bengio and Bengio [10] used MLP and Mixture of Gaussians based gating networks in a large-scale classification model. However, their approach requires the gating and experts networks to be retrained several times. Initially, the data is divided in random subsets, which are used to train the expert networks. The expert classifiers are used to determine an objective function for the gating network retraining. The new gating network then defines new subsets and the process is repeated until the termination criterion is fulfilled. This method requires the gating to be retrained on each iteration, hence making the procedure very time consuming. The SGA-II gating uses a fast sequential clustering in both stages, which, despite the simple structure, results in a high accuracy gating, as shown in section 4.

3.2 Non-redundant support vectors

Support Vector Machines are well-known for being a high computational complexity method on the recognition phase, specially for problems with high number of features or complex multiclass problems with high number of classes, as the branch networks in CombNET-III.

```
for $k \in \{1 \ldots K\}$
    Calculate $\bar{\mathbf{x}}_k = \ell_k^{-1} \sum_{i:y_i=k} x_i$
end for
Apply $SGA$ to $\{\bar{\mathbf{x}}_1 \ldots \bar{\mathbf{x}}_K\}$ obtaining $\{\nu_1 \ldots \nu_R\}$
for $j \in \{1 \ldots R\}$
    Apply $SGA$ to $\bigcup_{i:y_i=k, \bar{\mathbf{x}}_k \in \nu_j} x_i$
    obtaining $\left\{\varsigma_{j,1} \ldots \varsigma_{j,S_j}\right\}$
end for
```

Fig. 4 Proposed Method Algorithm (SGA-II)

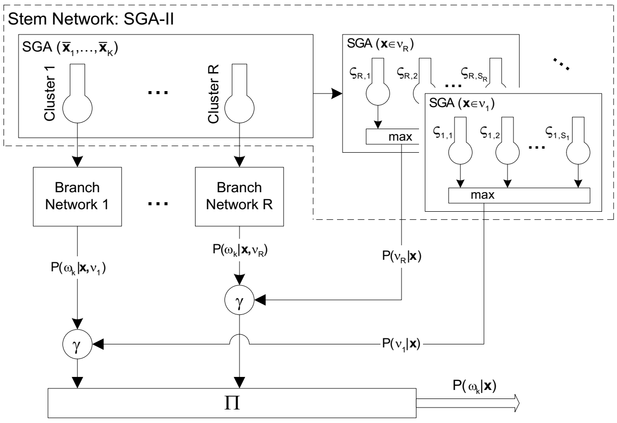

Fig. 5 CombNET-III with the two-layered self growing algorithm SGA-II

In previous works, several strategies based on the elimination of classifiers on the decoding phase had been introduced [17]–[19]. These methods, however, usually presents a performance penalty and make difficult to estimate the posterior probability of classes which classifiers had been eliminated along the decoding.

This paper introduces another approach, based on the fact that, as each sample is used to train many binary classifiers, usually these classifiers will present some support vectors in common. When a new sample $\mathbf{x}$ is presented to the classifiers, the Kernel value $K\left(\mathbf{x}, \mathbf{z}\right)$ will be the same in all classifiers that share the support vector $\mathbf{z}$, so, it only needs to be computed once.

The classification computational complexity of the multiclass SVM is:

$$O\left(N \sum_{h=1}^{H} SV_h^T\right) \tag{5}$$

where SV_h^T represents the i^{th} classifier's total number of support vectors, H is the total number of classifiers and N is the number of features. If the kernel value of $\mathbf{x}$ and the training samples that are support vectors in at least one classifier are calculated in advance, the complexity becomes:

$$O\left(N \cdot SV^{NR} + \sum_{h=1}^{H} SV_h^T\right) \tag{6}$$

where SV^{NR} is the number of non-redundant support vectors. Its is clear that $SV^{NR} \leq \sum_{h=1}^{H} SV_h^T$. The experimental results show that for most of the cases, including the experiments shown on this paper, $SV^{NR} \ll \sum_{h=1}^{H} SV_h^T$, reducing considerably the decoding computational complexity.

3.3 High confidence branch networks selection

Equation (2) uses all branch network results to generate

```
Calculate $P(\nu_j|\mathbf{x})$, $j \in \{1 \dots R\}$
Sort $\{\nu_1 \dots \nu_R\}$ so that:
    $P(\nu_j|\mathbf{x}) > P(\nu_{j+1}|\mathbf{x}) \forall j \in \{1 \dots R\}$
for $j \in \{1 \dots R\}$
    if $j > G$
        Set $P(\nu_j|\mathbf{x}) = 0$
        Set $P(\omega_k|\mathbf{x}, \nu_j) = 0.5$
    else
        Calculate $P(\omega_k|\mathbf{x}, \nu_j)$
    end if
end for
Calculate $P(\omega_k|\mathbf{x})$
```

Fig. 6 Branch networks selection algorithm

an output. In most cases, part of these outputs correspond to very low values that do not influence the final probability. Some previous applications of CombNET-II in embedded systems for handwritten digits recognition [15] used only the branch corresponding to the clusters with highest score on the stem network, reducing significantly the computational complexity. However, when applied to large-scale problems with large number of categories, this approach tends to compromise the accuracy, as the classification becomes more dependent on the gating network, which presents a low accuracy.

The SGA-II, however, presents a much higher accuracy than the original SGA used on previous works. Hence, branch networks corresponding to the lowest scores on the gating network can be eliminated from equation (2) with a higher confidence. The approach used on the experiments of this paper was to choose a fixed number G, $(1 \leq G \leq R)$ of the highest gating network probabilities, although one could also define a probability threshold.

The algorithm for calculating the final output using this procedure is shown in Figure 6. After the clusters posterior probabilities are sorted, the lowest values are set to zero and the correspondent branch networks' outputs are set to the random hypothesis. Thus, only the outputs of the branch networks corresponding to the G highest cluster probabilities are calculated, significantly reducing the computational complexity. The value of G is set experimentally, depending of the system requirements of accuracy and complexity.

4. Experiments

Two databases were used on the experiments, *Kanji400* and *Forest*. The *Kanji400* database, already used in [13], illustrates the efficiency of the proposed model in problems with large number of categories. The *Forest* database experiments explores the proposed model's behavior in a problem with large number of samples and very unbalanced categories.

4.1 ETL9B *Kanji400* database

This database consists of a subset of the ETL9B database[†]. The ETL9B database contains 3036 categories, composed by 2965 Chinese characters (Kanji) and 71 Japanese Hiragana characters. The first 400 classes were used, each contains 200 samples, from which 150 samples were used as the training set and 50 samples as the test set. The characters were resized by their largest dimension and the peripheral direction contributivity (PDC) feature extraction method [20] was applied.

Five different configurations of the gating network were tested, making the number of clusters in which the problem was divided equal to 5, 8, 12, 16 and 20. The classification accuracy for those configurations is shown in Figure 7, in which the circles' dotted line corresponds to the standard SGA-I algorithm recognition rate and the squares' solid line to the proposed SGA-II algorithm. The SGA-II subclusters were created from the same clusters generated in SGA-I. Moreover, the same SGA-I was used in CombNET-II and the original CombNET-III. The used similarity measurement was the normalized dot-product (the cosine between two vectors) and the inner potential threshold Θ_p is shown in Figure 7 under the x-axis (for details about the SGA algorithm, see [12], [13]). Even though the use of the similarity threshold Θ_s can speed up the convergence, it can also generate clusters with very few samples, deteriorating the balance among the clusters. For some large scale classification problems, this parameter can be used as a fine adjustment of the clusters' balance. On the *Kanji400* experiments, however, such adjustment is not necessary and Θ_s was set to -1 for all cases.

The number of sub-clusters on the SGA-II was chosen in order to keep its accuracy higher than 98%. As each data split generates different boundaries, which complexity depend on the size of the clusters and which categories they contain, the average amount of reference vectors per clusters is not proportional to the number of clusters. Although the complexity of SGA-II is higher than SGA-I, it is still much smaller than the branch networks and this increase can be neglected. For increasing number of clusters, the SGA-I presents a rapid decay on accuracy, as the linear hyperplanes between the clusters start to be responsible for more and more classes split, which true boundaries are usually very nonlinear. The multiple reference vectors of SGA-II make a better representation of those hyperplanes, achieving a significant increase on the gating network accuracy, specially for higher number of clusters.

For the CombNET-II experiments, the MLP neu-

[†] Available under request from
http://www.is.aist.go.jp/etlcdb

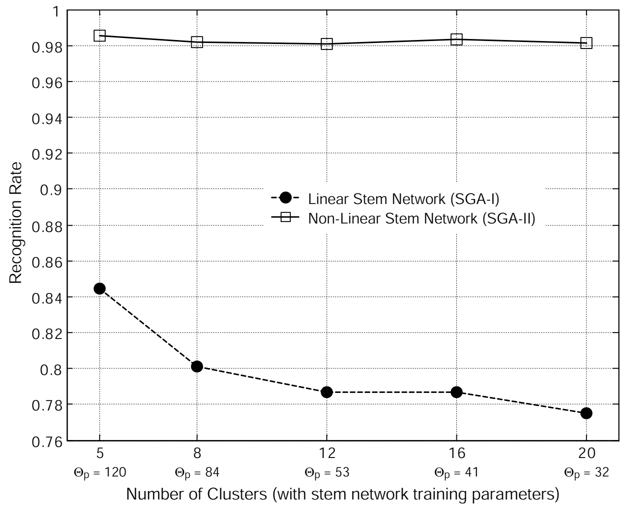

Fig. 7 Stem networks recognition rate results for the *Kanji400* database

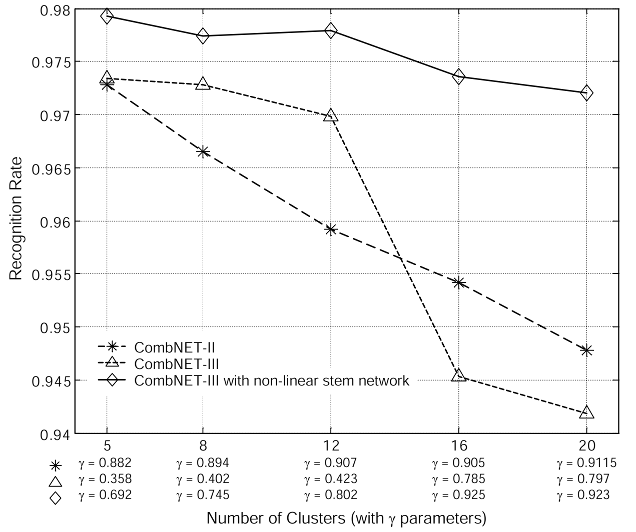

Fig. 8 Final recognition rate results for the *Kanji400* database

ral networks were trained by gradient descent backpropagation until the error was smaller than 10^{-4} or the iteration number exceeds 500, with learning rate equal to 0.1, momentum 0.9 and sigmoidal activation function slope 0.1. The number of hidden neurons and the γ parameter were optimized (by testing several values) for each experiment realization. In the case of CombNET-III (with both gating networks configuration), the binary SVM classifiers had non-biased output and a Gaussian kernel function, whose parameter σ was optimized for each experiment realization. The soft-margin C parameter was fixed at 200 (as several experimented values did not produce significant changes for the used data). For CombNET-III, each branch network training data was normalized to zero mean and unitary standard deviation.

Figure 8 shows the final recognition rate for the three models. The proposed model outperformed the other methods, achieving an error rate reduction between 16.8% and 51.9% in comparison with the original CombNET-III. It must be pointed that both the original CombNET-III and the proposed model used the same branch networks. CombNET-II shows an almost linear decreasing accuracy with increasing number of clusters. The original CombNET-III presents a better accuracy for small number of clusters, but also shows a rapid decrease for too many clusters. The proposed model presented a decrease of less than 1% from 5 to 20 clusters.

Previous works on CombNET-II showed that, for problems with large number of categories where each category belongs to only one clusters, the selection of few high confident branch networks results in a significant decrease in performance. Figure 9 shows the final recognition rate of CombNET-III with SGA-I and SGA-II for an decreasing number of computed branch networks. The x-axis represents the rate of considered branch networks for each number of clusters and the y-axis the proportional performance decrease, with 1.0 corresponding to the result when all branch networks are used. The dotted lines and solid lines corresponds to the SGA-I and SGA-II respectively. With SGA-II, there was no decrease until 50% and an almost negligible accuracy decrease until 20%. Using SGA-I, the result decreases from just one eliminated branch network and presents a rapid decline after 50%.

The final computational complexities of CombNET-III and the proposed modifications are shown in Figure 10. The circles' dotted line represents the original CombNET-III complexity for an increasing number of clusters. The diamonds' dashed line and squares' dotted line shows, respectively, the complexity when using the non-redundant support vectors strategy and the branch network reduction. For the later, the complexity is related to the smallest number of branch networks that presents the no accuracy reduction. If some tolerance is given, this complexity could be even smaller. Finally, the triangles' solid line shows the complexity of the complete proposed model, using both strategies. It is to be noticed that the y-axis is in logarithmic scale.

Table 1 described how these complexities were calculated, in which N is the number of features, R is the number of clusters on the case of divide-and-conquer methods, G' is the smallest group of branch networks that presents no decrease on performance, H_j is the number of binary SVM on the j^{th}, and SV_j^{NR} and SV_nh^T are, respectively, the number of non-redundant support vectors in the j^{th} multiclass and the number of support vectors in the h^{th} binary SVM of the j^{th} cluster.

The proposed model presents a computational complexity more than one order of magnitude smaller than the original CombNET-III. Again, the gating net-

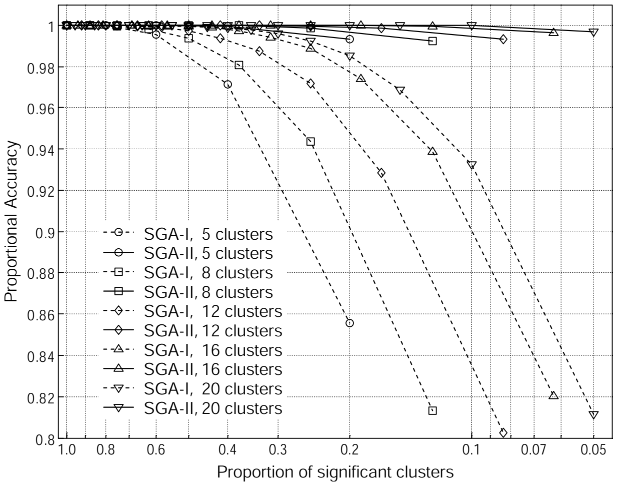

Fig. 9 High confidence branch networks selection recognition rate results for the *Kanji400* database

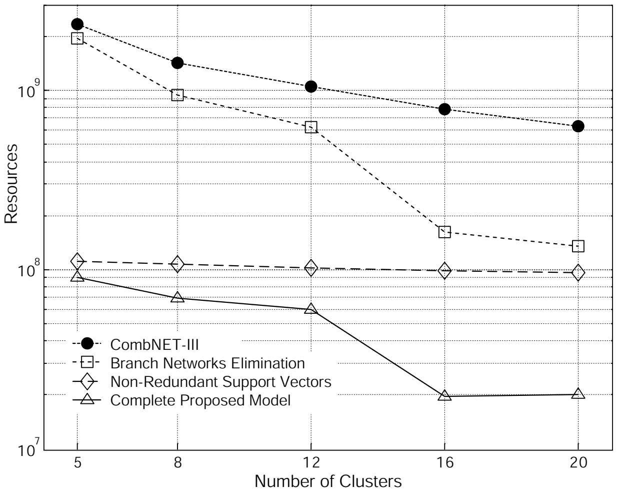

Fig. 10 Computational complexity for the *Kanji400* database

work complexity is not included on the equations of Table 1 as it is much smaller than the branch networks complexity.

Table 1 Classifiers computational complexity description

Classifier	Complexity Description
Original CombNET-III	$N \cdot \sum_{j=1}^{R} \sum_{h=1}^{H_j} SV_{jh}^T$
Non-Redundant Support Vectors	$\sum_{j=1}^{R} \left(N \cdot SV_j^{NR} + \sum_{h=1}^{H_j} SV_{jh}^T \right)$
High Confidence Branch Networks Selection	$N \cdot \sum_{j \in G'} \sum_{h=1}^{H_j} SV_{jh}^T$
Complete Proposed Model	$\sum_{j \in G'} \left(N \cdot SV_j^{NR} + \sum_{h=1}^{H_j} SV_{jh}^T \right)$

4.2 UCI KDD *Forest* database

This database, obtained from the UCI KDD Archive repository [21], consists of the forest cover type for 30 x 30 meter cells obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. It contains 581012 samples of 7 categories of forest cover, represented by 54 features, 10 quantitative values and 2 qualitative variables codified in 44 binary features. The first two classes, "SF" and "LP", represents more than 85% of the data, while the "WL"category contains only 0.47%, making this a very unbalanced problem. Table 2 shows the samples distribution, as well as how the data was split in three independent sets. Sequentially, for each 3 samples of each class, 2 were used for training and 1 for control/test. This second set was later split in two parts, again sequentially, with 1 sample for the control set and another for the test set.

The stem network was trained with raw data using Euclidean distance dissimilarity measurement, with parameters chosen in order to obtain 16 clusters. This is the minimal number of clusters generated by the SGA that produces branch networks which half-kernel matrixes fit on 3GB of memory (the largest one (ν_3) contains 35231 samples). Also, after the training, if a cluster contains samples of a class that represents less than 10% of the cluster, these samples are transferred to the nearest cluster that contains this class. This procedure helps to keep the balance inside each cluster, although some clusters still present some unbalance. For instance, ν_3 contains only 166 samples of class "PD" and 14565 samples of class "SF". This unbalance does not seriously affect the accuracy, but adds unnecessary complexity. The total number of subclusters generated by the SGA-II is 1400 (average of 87.5 per cluster).

As each class belongs to several clusters, it is not possible to calculate the accuracy of the stem network. In order to verify its performance, only the amount of control data set samples with highest score on each cluster was verified. This matched the clusters sizes with a difference up to 0.28% of the total number of samples in each data set.

Each branch network parameters were optimized independently by the accuracy of the control data set. However, as it is not possible to define which samples of the control data set should be used for optimizing

Table 2 *Forest* database samples distribution and data sets

Class	Training	Control	Test	Total	Rate
SF	141227	35306	35307	211840	36.46%
LP	188867	47217	47217	283301	48.76%
PD	23836	5959	5959	35754	6.15%
WL	1832	458	457	2747	0.47%
AP	6328	1582	1583	9493	1.63%
DF	11578	2895	2894	17367	2.99%
KH	13676	3415	3419	20510	3.53%
Total	387344	96832	96836	581012	100.0%

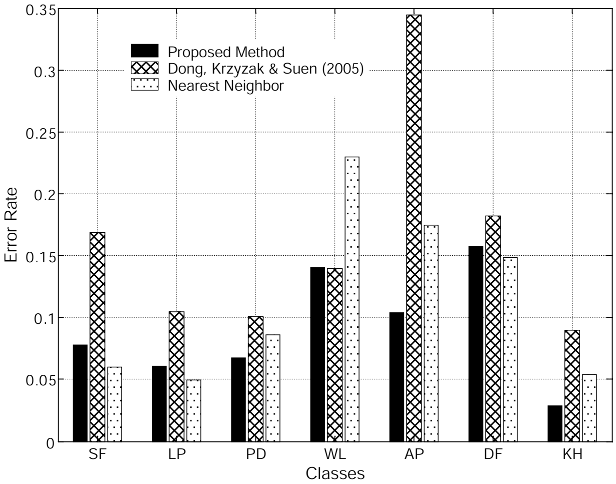

Fig. 11 Individual class error rates for the *Forest* database

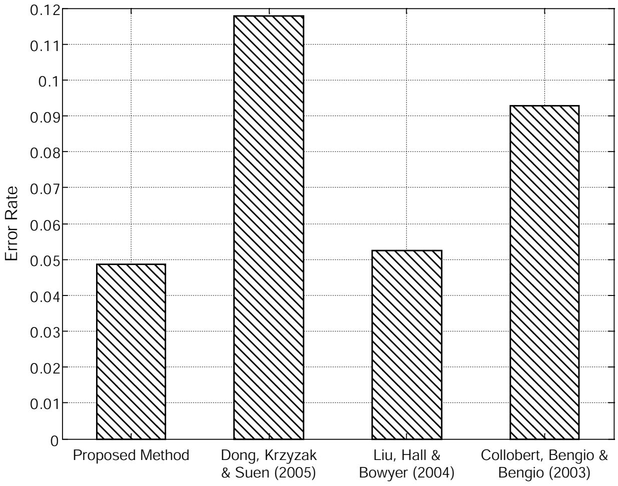

Fig. 12 Error rates for the *Forest* binary classification problem

each branch network, the gating network probability was used define these splits. For instance, given a sample $\mathbf{x}$, if $P(\nu_i|\mathbf{x}) = \max_j P(\nu_j|\mathbf{x})$, the sample $\mathbf{x}$ will be used to optimize the i^{th} branch network. The average accuracy for all classes in all clusters achieved 90.36%. The γ parameter (from equation (2)) and G (described in section 3.3) were optimized based on the control data set average accuracy of all classes, being respectively 0.153 and 2.

The individual classes error rates for the test data are shown in Figure 11. Due to the use of different data splits, it is difficult to make comparisons with other authors' results. Nevertheless, Figure 11 also includes the results presented by Dong, Krzyzak and Suen [22]. They used a similar splitting of data (75% for training and 25% for testing), with a One-versus-Rest single multiclass SVM trained by decomposition. Furthermore, Figure 11 includes the result for the k-Nearest Neighbor (kNN) classifier, which parameters $K = 1$ was found by the accuracy in the control data. Figure 11 does not include the results for the original CombNET-III. When the SGA-I gating is used on this database, the control samples for each branch network cannot be properly selected and the SVM parameters cannot be optimized. Also, the real gating accuracy is probably very low, due to the high number of clusters.

The proposed method outperformed both compared methods. The final averaged error for all classes was 9.072%. For the method described in [22], the averaged error was 16.144%. Surprisingly, the kNN method performed better than the method from [22], resulting in an averaged error of 11.460%. On reason for this can be the method used by Dong, Krzyzak and Suen for splitting the data. The *Forest* database samples for the larger categories presents a large variation, with samples on the beginning of the original file being very different from the ones on its end.

Liu, Hall and Bowyer [23] and Collobert, Bengio and Bengio [10] considered only the binary classification of class "SF" against the others, using respectively an ensemble of decision trees and a mixture of SVMs. They selected a training data set of 100000 samples and a test set of 50000. Both the proposed model and the model from Dong, Krzyzak and Suen were not trained specifically for this binary classification problem. Nevertheless, considering that misclassifications between classes different from "SF" are not errors, the results can be compared. The γ parameter was optimized for this purpose, obtaining 0.300. Figure 12 presents the results for this binary classification task, comparing the proposed model with the results from [10], [22], [23].

The proposed model obtained the best accuracy. Of course, this is not a proper comparison, as different data splits were used, and the experiment's objective are different. Nevertheless, it illustrates the flexibility of CombNET-III with the SGA-II gating network. CombNET-II presented good results on unbalanced classification problems [24], and the results obtained by the proposed model, which does not use redundant training samples among the branches, are encouraging.

5. Discussion and Conclusions

This paper proposed an extension of the large-scale classification model CombNET-III. The main objectives of this extension were to improve the accuracy over the original CombNET-III and to reduce the classification computational complexity. The main proposed modification was the use of a nonlinear gating network, named SGA-II, which represents each cluster by several reference vectors, achieving higher accuracy. This higher accuracy permits the elimination of the less confident branch networks, reducing the computational complexity. Moreover, a new strategy for reducing the number of Kernel function calls in each multiclass SVM branch network was presented. The use of a more accurate gating network also enables the application of the proposed model in problems with large number of samples.

The use of SGA-II proportionated a significant accuracy improvement for the *Kanji400* database, with small complexity increase. This enables the use of a larger number of clusters, reducing the complexity. Moreover, the high-confidence branch networks selection was shown to be efficient with the use of SGA-II. In some cases, more than half of the branches could be ignored with no accuracy penalty. The non-redundant support vectors strategy, although asking for a more complex implementation, reduced the complexity by more than one order of magnitude, being an efficient alternative for speeding-up multiclass SVM classification.

The results on the *Forest* database shows that CombNET-III with SGA-II is an important alternative not only for problems with large number of categories, but also for problems with large number of samples and/or unbalanced problems. By splitting the data, "less unbalanced" smaller problems can be efficiently solved.

Even though an important reduction on computational complexity was achieved, further investigation about feature subset selection on the branch networks could reduce this complexity even more. Refinements of the branch network selection procedure (e.g. the use of a probability threshold) also need to be explored. A deeper investigation of the application of the proposed model on unbalanced problems and also in problems with a much higher number of categories is also necessary. Another important research direction is the use of more sophisticated training procedures on the branch networks for reducing training time.

Acknowledgments

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. The research of the third author is partially supported by the Grant-in-Aid for Private University High-Tech Research Center from Ministry of Education, Culture, Sports, Science and Technology, Government of Japan.

References

[1] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, New Jersey, 2004.

[2] R.A. Jacobs, M.I. Jordan, G.E. Hinton, and S.J. Nowlan, "Adaptive mixtures of local experts," Neural Computation, vol.3, no.1, pp.79–87, 1991.

[3] M. Arai, J. Wang, K. Okuda, and J. Miyamichi, "Thousands of hand-written kanji recognition by "HoneycombNET"," IEICE Transactions on Information & Systems, vol.J76-D-II, no.11, pp.2316–2323, November 1993.

[4] M. Arai, K. Okuda, and J. Miyamichi, "Thousands of handwritten kanji recognition by "HoneycombNET-II"," IEICE Transactions on Information & Systems, vol.J77-D-II, no.9, pp.1708–1715, September 1994.

[5] M. Arai, K. Okuda, H. Watanabe, and J. Miyamichi, "A large scale neural network "HoneycombNET-III" that has a capability of additional learning," IEICE Transactions on Information & Systems, vol.J80-D-II, no.7, pp.1955–1963, July 1997.

[6] K. Saruta, N. Kato, M. Abe, and Y. Nemoto, "A fine classification method of handwritten character recognition using exclusive learning neural network (ELNET)," IEICE Transactions on Information & Systems, vol.J79-D-II, no.5, pp.851–859, May 1996.

[7] K. Saruta, N. Kato, M. Abe, and Y. Nemoto, "High accuracy recognition of ETL9B using exclusive learning neural network - II (ELNET-II)," IEICE Transactions on Information & Systems, vol.E79-D, no.5, pp.516–522, May 1996.

[8] J.T.Y. Kwok, "Support vector mixture for classification and regression problems," Proceedings of the International Conference on Pattern Recognition (ICPR'98), Brisbane, Queensland, Australia, pp.255–258, 1998.

[9] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," Neural Computation, vol.14, no.5, pp.1105–1114, May 2002.

[10] R. Collobert, S. Bengio, and Y. Bengio, "Scaling large learning problems with hard parallel mixtures," International Journal on Pattern Recognition and Artificial Intelligence, vol.17, no.3, pp.349–365, 2003.

[11] A. Iwata, T. Touma, H. Matsuo, and N. Suzumura, "Large scale 4 layered neural network "CombNET"," IEICE Transactions on Information & Systems, vol.J73-D-II, no.8, pp.1261–1267, August 1990.

[12] K. Hotta, A. Iwata, H. Matsuo, and N. Susumura, "Large scale neural network CombNET-II," IEICE Transactions on Information & Systems, vol.J75-D-II, no.3, pp.545–553, March 1992.

[13] M. Kugler, S. Kuroyanagi, A.S. Nugroho, and A. Iwata, "CombNET-III: a support vector machine based large scale classifier with probabilistic framework," IEICE Transactions on Information & Systems, vol.E89-D, no.9, pp.2533–2541, September 2006.

[14] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, ed. A.J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, pp.61–74, MIT Press, Cambridge, MA, March 1999.

[15] 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," Proceedings of the 4th International Conference on Engineering Application of Neural Networks, Gibraltar, pp.40–46, June 1998.

[16] Y. Waizumi, N. Kato, K. Saruta, and Y. Nemoto, "High speed and high accuracy rough classification for handwritten characters using hierarchical learning vector quantization," IEICE Transactions on Information & Systems, vol.E83-D, no.6, pp.1282–1290, June 2000.

[17] J.C. Platt, N. Cristianini, and J. Shawa-Taylor, "Large margin DAGs for multiclass classification," Advances in Neural Information Processing Systems, vol.12, pp.547–553, 2000.

[18] B. Kijsirikul, N. Ussivakul, and S. Meknavin, "Adaptive directed acyclic graphs for multiclass classification," Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence, pp.158–168, Springer-Verlag, 2002.

[19] B. Kijsirikul, N. Boonsirisumpun, and Y. Limpiyakorn,

"Multiclass support vector machines using balanced dichotomization," Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence, ed. C. Zhang, H.W. Guesgen, and W.K. Yeap, LNAI 3157, Berlin, pp.973–974, Springer-Verlag, August 2004.

[20] N. Hagita, S. Naito, and I. Masuda, "Chinese character recognition by peripheral direction contributivity feature," IEICE Transactions on Information & Systems, vol.J66-D, no.10, pp.1185–1192, October 1983.

[21] S. Hettich and S.D. Bay, "The UCI KDD Archive." Irvine, CA: University of California, Department of Information and Computer Science, 1999. http://kdd.ics.uci.edu.

[22] J. Dong, A. Krzyzak, and C.Y. Suen, "Fast svm training algorithm with decomposition on very large data sets," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no.4, pp.603–618, April 2005.

[23] X. Liu, L.O. Hall, and K.W. Bowyer, "Comments on "A parallel mixture of SVMs for very large scale problems"," Neural Computation, vol.16, no.7, pp.1345–1351, July 2004.

[24] A.S. Nugroho, S. Kuroyanagi, and A. Iwata, "A solution for imbalanced training sets problem by CombNET-II and its application on fog forecasting," IEICE Transactions on Information & Systems, vol.E85-D, no.7, pp.1165–1174, July 2002.

Mauricio Kugler received the degree in electrical engineering in 2000, and the MSc degree in biomedical engineering in 2003, both from the Federal Technological University of Parana, Brazil. In 2007, he received a PhD degree in computer science and engineering from the Nagoya Institute of Technology, Japan. Currently, he is an assistant professor at the Department of Computer Science and Engineering at this same institute. His research interests include machine learning, large scale pattern recognition methods, biomedical signals processing, spiking neural networks and hardware programming. He is a member of the Institute of Electrical & Electronics Engineers (IEEE).

Susumu Kuroyanagi received a B.S. in 1991 from the Department of Electrical and Computer Engineering at the Nagoya Institute of Technology. He completed the first half of the doctoral program in 1993 and the second half in 1996, receiving the D.Eng. degree from the same institute. In 1996, he became a research associate in the Department of Electrical and Computer Engineering at the Nagoya Institute of Technology, and, in 2003, a research associate in the Graduate School of Engineering, at the Department of Computer Science and Engineering. Since 2006, he has been an associate professor in this same Graduate School. He is engaged in researches about neural networks and auditory information processing, also being a member of the Acoustic Society of Japan, the Japan Neural Network Society and Japanese Society for Medical and Biological Engineering.

Anto Satriyo Nugroho is a researcher working for Agency for the Assessment & Application of Technology (BPPT), Indonesia. He received his B.Eng degree in 1995, M.Eng in 2000 and Dr.Eng degree in 2003, all in Electrical & Computer Engineering from Nagoya Institute of Technology, Japan. From 2003 to 2007, he is working for Chukyo University as visiting professor in School of Life System Science & Technology. His research interest is in the field of pattern recognition, bioinformatics and data mining. He is a member of the Institute of Electrical & Electronics Engineers (IEEE).

Akira Iwata received a B.S. in 1973 from the Department of Electrical Engineering, Faculty of Engineering, Nagoya University. He completed the M.E. program in 1975 and became a research associate in the Department of Information, Nagoya Institute of technology. He was a visiting researcher from April 1982 to October 1983 in the research Institute of Medical information, University of Giessen Medical School, Germany. He became an associate professor in the Department of Information, Nagoya Institute of Technology in 1984, and a professor in the Department of Electrical and Computer Engineering in 1993, and vice president in 2002, and has been a professor in the Department of Computer Science and Engineering, Graduate School, since 2004. He is engaged in research on neural networks and internet security, He holds a D.Eng. degree. He received an IEICEJ paper Award in 1993 and an Information Processing Society Best Author Award in 1998. He is a member of the Information Processing Society, JSMEBE, the Japan Electrocardiography Society, the Japan Neural Network Society, and Japan Society for Medical Information Processing. He is an IEEE Senior Member.