FUZZY INTEGRAL BASED-MIXTURE TO SPEED UP THE ONE-AGAINST-ALL MULTICLASS SVMS

H. Nemmour and Y. Chibani

Signal Processing Laboratory, Faculty of Electronic and Computer Sciences
University of Sciences and Technology Houari Boumediene
EL-Alia B. P. 32, 16111, Algiers, Algeria

ABSTRACT

The One-Against-All (OAA) is the most widely used implementation of multiclass SVM. For a K-class problem, it performs K binary SVMs designed to separate a class from all the others. All SVMs are performed over the full database which is, however, a time-consuming task especially for large scale problems. To overcome this limitation, we propose a mixture scheme to speed-up the training of OAA. Thus, each binary problem is divided into a set of sub-problems trained by different SVM modules whose outputs are subsequently combined throughout a gating network. The proposed mixture scheme is based on Sugeno's fuzzy integral in which the gater is expressed by fuzzy measures. Experiments were conducted on two benchmark databases which concern Handwritten Digit Recognition (ODR) and Face Recognition (FR). The results indicate that the proposed scheme allows a significant training and testing time improvement. In addition, it can be easily implemented in parallel.

1. INTRODUCTION

Due to their strong learning ability and high generalization performance compared to conventional classifiers such as Naïve Bayes and artificial neural networks, Support Vector Machines (SVMs) are used in a wide range of applications. However, their major limitation is that the learning time is at least quadratic to the number of training data. Thereby, for large scale problems such as face and handwriting recognition where data have high dimensions, the training of SVMs is a time-consuming task. Specifically, the training complexity increases if several classes are included. Since SVMs are originally formulated for binary classification, their extension to multiclass problems is not unique. The earliest and the most commonly used implementation for multiclass SVM is the One-Against-All (OAA) method. For a K-class problem, it constructs K SVMs trained respectively, to separate a class from all the others. Each SVM is performed over the full database and the training time is proportional to $K \times T^2$, where T is the number of training samples. Thus, using the same number of data, the resolution of a multiclass SVM is computationally much more expensive than a binary SVM [8].

Many research works have been done over the past few years to speed up the SVM training. Interestingly, we notice the mixture scheme introduced in [7]. The underlying idea consists of using different experts in different regions of input space. Thus, the original problem is divided into a set of sub-problems easier to learn. Each sub-problem is trained by a particular SVM. Then, outputs of trained modules are combined by a gater which is generally an artificial neural network trained on the whole database [5]. To mitigate this issue, [6] proposed the use of local gaters trained only over their own sub-problems. However, the use of such gaters is not trivial in practice. First, many iterations are required to train the neural network, which will compromise the aim of mixture that is training time reduction. Besides, the user is faced to a large variety of learning rules and possible architectural selections. More recently, a Min-Max modular SVM approach was employed in [8] to speed up text categorization. Unfortunately, the modularity concept as well as experimental analysis were drawn independently for each class, which is different to the multiclass reasoning.

In the work presented here, we propose the use of mixture concept to speed up the OAA SVM method. Another contribution of this paper consists of using Sugeno's fuzzy integral as a mixture scheme in which local gaters are expressed by fuzzy measures. The main advantage of our approach is that there is no need to any training stage or parameter selection. Furthermore, this scheme was successfully used for mixtures of binary SVMs using the USPS handwriting task in [9]. Presently, experiments are conducted on two different applications which are ODR and face recognition using standard databases. In the next section, we briefly review SVMs for binary and multi-class problems. In addition, we introduce the fuzzy integral-based mixture. Experimental results are presented and discussed in section 3 while the last section gives the main conclusions of the paper.

2. THEORETICAL CONCEPTS

2.1. Support vector machines

SVMs construct binary classifiers from a set of training examples such that: $(x_j, y_j) \in R^N \times \{\pm 1\}, j = 1,\ldots,n$ [1, 10]. The data are mapped into a dot product space via a kernel function such that: $K(x,x') = \langle \phi(x), \phi(x') \rangle$. Then, the solution is expressed in terms of kernel expansion as:

$$f(x) = sign\left(\sum_{j=1}^{Sv} \alpha_j K(x,\, x_j) + b \right) \tag{1}$$

Sv is the number of support vectors which are training data for which $0 \le \alpha_i \le C$, while the bias *b* is a scalar. Note that *C* is the cost parameter. Furthermore, a large number of mathematical functions are eligible to be a SVM-kernel. Here, we use the Radial Basis Function kernel given in (2) (σ and *C* are user defined parameters).

$$K(x, x_j) = \exp\left(-\frac{1}{2\sigma^2} \|x - xj\|^2 \right) \tag{2}$$

The earlier extension of SVMs for multiclass problems is the One-Against-All (OAA) approach. For a *K*-class problem, *K* classifiers are performed to separate iteratively each class from all the others. Therefore, a sample x is assigned to the class with the maximum positive output.

$$\arg \max_{k}^{K} f_k(x) \tag{3}$$

2.2. Fuzzy integral-based mixture

The basic idea of mixture is to break a complex learning problem into a set of sub-problems each of which is trained by a SVM module [5, 7]. Then, module outputs are linearly combined by a gater according to:

$$y = \sum_{i=1}^{n} f_i(x) \cdot \pi(x) \tag{4}$$

$f_i(x)$ is the output of the i^{th} expert for a sample *x*. π is the gater output while *n* is the number of experts (SVMs). To be efficient, the gating network needs a long training stage with a carefully parameter selection. To overcome this limitation, we adapt the fuzzy integral-combination rule to the mixture concept. The fuzzy integral combines non linearly, objective evidences in the form of expert decisions with respect to subjective evaluation of their performance expressed by a fuzzy measure. Various forms of the fuzzy measure are available in literature. The most widely used is the so-called g_λ measure. Our approach can be summarized in four main steps. First, a two-class problem to separate a class *k* from non class *k* is divided into a set of small two-class problems each of which is trained by a SVM module. Then, SVM outputs are handled through a fuzzy model to express their membership degrees in positive classes. For each class, membership degrees derived from all SVMs are combined with respect to fuzzy measures to calculate the fuzzy integral. Finally, mixture responses are compared according to the conventional OAA decision function to classify data.

2.2.1. g_λ fuzzy measure

The g_λ fuzzy measure is defined to satisfy the following property [2, 3]: Let $Z = \{z_1,\ldots,z_n\}$ be the set of available experts. For each expert z_i to be combined, we associate a fuzzy measure $g_k(z_i)$ indicating its performance in the class *k*. For a given sample, let $h_k(z_i)$ be the objective evidence of the expert z_i in the class *k*. The set of experts is then rearranged such that the following relation holds: $h_k(z_1) \ge \cdots \ge h_k(z_n) \ge 0$. Thus, we obtain an ascending sequence of SVMs $A_i = \{z_1,\ldots,z_i\}$, whose fuzzy measures are constructed as:

$$\begin{aligned} g_k(A_1) &= g_k(z_1) \\ g_k(A_i) &= g_k(A_{i-1} \cup z_i) \\ &= g_k(A_{i-1}) + g_k(z_i) + \lambda g_k(A_{i-1}) g_k(z_i) \end{aligned} \tag{6}$$

For a class *k*, λ is the unique root of a *n*-1 degree equation that is $\lambda \in]-1,..,+\infty[$ and $\lambda \ne 0$. It is determined by solving the following equation:

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda\, g_k(z_i)) \tag{7}$$

The relation in (6) provides both the weight of a single expert and the weight of a subset of experts. Unfortunatly, there is no rule which would be followed to attribute g_k values. They can be subjectively assigned by the user, or generated from training data [2].

2.2.2. Fuzzy integral

The fuzzy integral I_S, of a function $h: Z \to [0,1]$ with respect to a fuzzy measure *g* over *Z* is computed by [2-4]:

$$I_S(k) = \underset{i=1}{\overset{n}{Max}} [Min(h_k(z_i),\, g_k(A_i))] \tag{8}$$

2.2.3. Fuzzy membership model

Once trained, the SVM output is either positive or negative. However, the fuzzy integral involves the use of objective evidences. We propose in what follows a fuzzy model to express membership degrees of SVM outputs in the positive classes (note that there is no need to evaluate the membership degrees in negative classes, since they cannot inform about the appurtenance of the object to a particular class). Let $f_{z_i}(x)$ be the output of the SVM z_i obtained for an object x to be classified. Its membership $h_+(z_i)$ in the positive class is defined as:

Table 1. Fuzzy class membership model

If $f_{z_i}(x) > 1$ *then* $h_+(z_i) = 1$
Else
If $(f_{z_i}(x) < -1)$ *then* $h_+(z_i) = 0$
Else
$$h_+(z_i) = \frac{1 + f_{z_i}(x)}{2}$$

According to this fuzzy model the decision functions of all positive classes calculated through the fuzzy integral are compared to classify data such that:

$$\arg\max_{k=1}^{K} I_{S_+}(k) \tag{9}$$

3. EXPERIMENTAL RESULTS

The validity of the proposed scheme is investigated on two different problems of pattern recognition which are Optical Digit Recognition (ODR) and Face Recognition (FR) using standard databases. ODR experiments are conducted on USPS handwriting recognition task. This database contains gray-level images of ten numeral classes (0-9), extracted from US postal envelopes. All images are normalized to a size of (16×16) pixels yielding 256 dimensional datum vector. This database contains a large number of data (9298 character), which is very difficult to train with the OAA. Thereby, we selected 1200 samples for the training stage and 300 samples for the test stage. Furthermore, the FR experiments are carried out using the Cambridge ORL database which contains 40 classes (each person corresponds to a distinct class). For each person, there are 10 images taken against a dark homogeneous background with different facial expressions such as open or close eyes, and smiling or not smiling, etc. The images have a size of (112×92) pixels yielding a 10304 dimensional input vector. For all classes, 5 images were randomly selected to be used in the training stage while the others were used in the test stage. For both applications, training and testing data are adjusted to [0, 1].

Before performing the mixture tests, C and σ were tuned experimentally using a subset of both databases. Several SVMs using different kernel and cost parameters: $\sigma = [1,\cdots,150]$ and $C=[1,\cdots,500]$, were performed. Then, the pair which achieves the smallest error rate was retained for the mixture tests. Thus, for OCR $(C,\sigma)=(500,10)$ and for FR $(C,\sigma)=(100,150)$.

Furthermore, we have shown in [9] that for binary SVMs, the fuzzy integral-based mixture allows a Training Time Reduction (TTR) which reaches 82.79% by using only 2 modules. Presently, we attempt to use mixtures of 2 and 3 modules to reduce the Training Time (TT) as well as the Recognition Time (RT) of the OAA method without reducing the Overall Recognition Rate (ORR). The fuzzy measures are assigned by calculating the ORR of the trained modules over the whole database. The results obtained for ODR and FR databases are displayed in figures 1 and 2, respectively. These figures depict in addition to the ORR, the Relative Training and Recognition Times, namely, RTT and RRT calculated as:

$$RTT_i = \frac{TT_i}{TT_1}\ ,\ \ RRT_i = \frac{RT_i}{RT_1}$$

$i = 2,3,4$ and designates the number of modules in the mixture while $i = 1$ designates the SVM trained over the full database.

As can be seen, for both problems, the training and the recognition speeds grow rapidly according to the number of modules. Specifically, for ODR using 2 modules the RTT and the RRT are about 53% and 36.36% with respect to the times required to solve the whole problem, and are less than 30% with 3 modules. Moreover, there is a clear performance improvement in ORR compared to the single module. Similarly, the results obtained for the FR problem show that the RTT and RRT decrease rapidly using 2 and 3 modules while keeping the same performance as the SVM trained over the whole database. However, above 3 modules, the TTR and the RRT decrease slightly whereas the ORR is bed. This outcome is related to the fact that when the modules become very small compared to the whole database, they have weak generalization abilities (small g values) since they do not contain enough training data. This leads systematically to an accuracy lower than that of the SVM trained on the original problem.

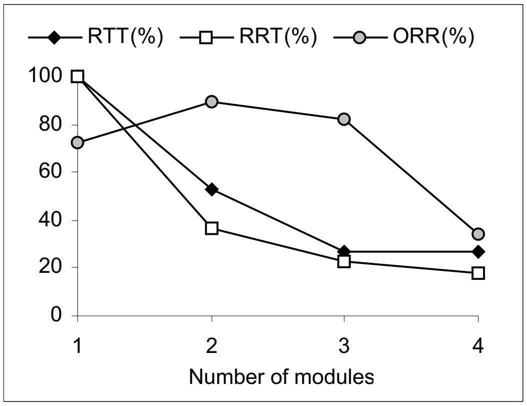

Fig 1. Mixture performance against the single SVM performance (ODR database)

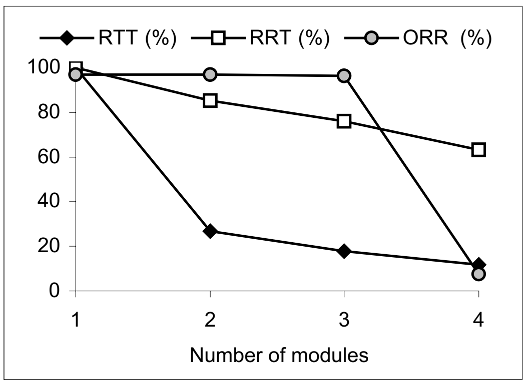

Fig 2. Mixture performance against the single SVM performance (FR database)

4. CONCLUDING REMARKS

We have presented a new mixture for SVMs which was employed to speed up the OAA multiclass SVM method. In this approach, a problem designed to separate a class from all other classes is divided into a set of sub-problems trained separately and combined by using the fuzzy integral. Trained modules produce complementary information about the data included in the gater which is expressed by fuzzy measures generated from the original problem. Moreover, the main advantage of this mixture scheme is that the gater does not need a training stage in contrast to conventional gaters such as neural networks. Experiments were conducted on two pattern recognition applications containing different numbers of classes (10 classes for ODR and 40 for FR). The results obtained have shown that for both applications, the division of the original problem into 2 and 3 modules, the mixture gives a significant improvement in training and testing speeds while keeping at least the performance of the SVM trained over the full database. Furthermore, this scheme can be easily parallelized to be again much more faster since it uses independent sub-problems.

5. REFERENCES

[1] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Edited by Ussama Fayyad, vol. 2, pp. 121-167, 1998.

[2] S. B. Cho, and J. H. Kim, "Combining Multiple Neural Networks by Fuzzy Integrals for Robust Classification," *IEEE Trans. Systems, Man and Cybernetics*, vol. 25, pp. 380- 384, 1995.

[3] S. B. Cho, "Fuzzy Aggregation of Modular Neural Networks With Ordered Weighted Averaging Operators," *International journal of approximate reasoning*, vol.13, pp. 359-375, 1995.

[4] S. B. Cho, "Fusion of Neural Networks with Fuzzy Logic and Genetic Algorithm," *Integrated Computer-Aided Engineering*, *IOS Press*, vol. 9, pp. 363–372, 2002.

[5] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," *Neural Computation*, vol. 14, pp. 1105-1114, 2002.

[6] R. Collobert, Y. Bengio, and S. Bengio, "Scaling large learning problems with hard parallel mixture," *International Journal of pattern recognition and artificial intelligence*, vol. 17, pp. 349-365, 2003.

[7] J.T.Y. Kwok, "Support vector mixture for classification and regression problems," *Proceedings of the 14th ICPR International Conference on Pattern Recognition*, Brisbane, Australia, August, 1998, pp. 255-258.

[8] F. Y. Liu, K.We, H. Zhao, and B. L. Lu, "Fast Text Categorization with Min-Max Modular Support Vector Machines", *In Proc. International Joint Conference on Neural Networks*, Montreal, Canada, July 31- August 4, 2005, pp.570-575.

[9] H. Nemmour and Y. Chibani, "Fuzzy integral for a rapid mixture of support vector machines", In Proc. *International Joint Conference on Neural Networks*, Montreal, Canada, 29 July-4 August 2005, pp. 901-906.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Verlag, 1995.