LETTER

Support Vector Machines with Binary Tree Architecture for Multi-Class Classification

Sungmoon Cheong
Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
E-mail : insomnia@etri.re.kr

Sang Hoon Oh
Department of Electrical Engineering, University of Mokwon
Daejeon, Korea
E-mail : shoh@mokwon.ac.kr

Soo-Young Lee
Department of BioSystems, Department of Electrical Engineering and Computer Science,
and Brain Science Research Center, Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
E-mail : sylee@kaist.ac.kr

(Submitted on January 21, 2004)

Abstract**—** For multi-class classification with Support Vector Machines (SVMs) a binary decision tree architecture is proposed for computational efficiency. The proposed SVM-based binary tree takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs. A modified Self-Organizing Map (SOM), K-SOM (Kernel-based SOM), is introduced to convert the multi-class problems into binary trees, in which the binary decisions are made by SVMs. For consistency between the SOM and SVM the K-SOM utilizes distance measures at the kernel space, not at the input space. Also, by allowing overlaps in the binary decision tree, it overcomes the performance degradation of the tree structure, and shows classification accuracy comparable to those of the popular multi-class SVM approaches with "one-to-one" and "one-to-the others".

Keywords**—**Support Vector Machine, multi-class classification, Self-Organizing Map, binary decision tree

1. Introduction

Among many classification methods the Support Vector Machine (SVM) had demonstrated superior performance. [1-6] However, the SVM was originally developed for binary decision problems, and its extension to multi-class problems is not straightforward. The popular methods for applying SVMs to multi-class classification problems decompose the multi-class problems into many binary-class problems and incorporate many binary-class SVMs. [7-9] For examples, an N-class problem needs $N(N-1)/2$ binary SVMs with the "one-against-one" approach, while N SVMS for the "one-against-the others" approach. Although the "one-against-one" approach demonstrates superior performance, it may require prohibitively-expensive computing resources for many real-world problems. The "one-to-the others" approach shows somewhat less accuracy, but still demands heavy computing resources, especially for real-time applications.

In this Letter we propose a hierarchical binary decision tree, of which binary decisions are made by SVMs. The proposed hierarchical SVM, SVM-BTA (Support Vector Machines with Binary Tree Architecture), takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs. Although $(N-1)$ SVMs need be trained for an N-class problem, it only requires to test $log_2 N$ SVMs for the classification decision. In practice the overlap in the tree-structured decision regions may be allowed to reduce performance degradation with slightly more SVMs. The essence of this Letter is an algorithm to convert the multi-class problems into binary-tree architectures without much performance degradation, and the Kernel-based Self-Organized Map (K-SOM) is introduced for this purpose.

In Section 2 we describe the basic architecture of the proposed SVM-BTA (Support Vector Machines with Binary Tree Architecture). The Kernel-based SOM (K-SOM) is introduced to convert the multi-class problems into SVM-based binary-tree architectures in Section 3. Experimental results are presented to compare the performance of the proposed SVM-BTA with traditional multi-category approaches in Section 4.

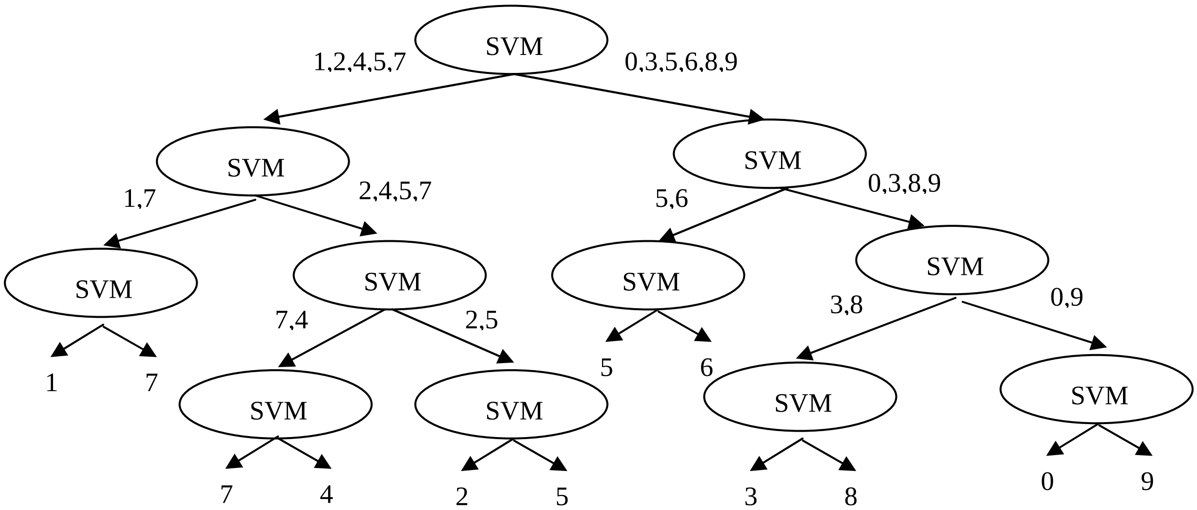

Figure 1. Support Vector Machines with Binary Tree Architecture (SVM-BTA) for the classification of 10 classes from “0” to “9”. The overlaps are allowed for the classes “5” and “7”.

2. Support Vector Machines with Binary Tree Architecture (SVM-BTA)

As shown in Figure 1, the SVM-BTA (Support Vector Machines with Binary Tree Architecture) solves an *N*-class pattern recognition problem with a hierarchical binary tree, of which each node makes binary decisions with an SVM. The hierarchy of binary decision subtasks should be carefully designed before the training of each SVM classifier.

In this Letter we propose 2 methods to convert the multi-class problems into SVM-BTA. In the first method the results of the K-SOM is plotted in 2-dimensional output space, and a human draws a linear line to divide the classes into 2 groups with minimum confusion. The second method utilizes automatic grouping, which maximizes the scattering parameter defined as [10]

$$S = \frac{\| \mathbf{m}_1 - \mathbf{m}_2 \|^2}{s_1^2 + s_2^2} . \tag{1}$$

It is equivalent to maximize the distance between two group centers and minimize the variance in each group. For each of the 2^N grouping possibilites with *N*-classes the scattering measures are calculated and the one with the highest scattering measure is selected. Then, the grouping at the lower nodes is made similarly.

To reduce the performance degradation due to improper grouping we also allow overlaps between the two groups. Since the misclassification tends to occur with confusing samples on the decision boundary, the overlap actually provides a chance to correct those confusing samples and results in better performance.

3. Kernel-based Self-Organizing Map (K-SOM)

At each node of the binary tree the decision is made to assign the input pattern into one of two groups, which consist of multiple classes. There exist many ways to divide the multi-classes into 2 groups, and it is critical to have proper grouping for the good performance of SVM-BTA. We propose to use the Self-Organizing Map (SOM), which maintains the relationship among input patterns, to convert the input space into 2-dimensional space for easy visualization. However, the standard SOM utilizes Euclidean distance in the input space for the relationship, while the SVM classifies in the kernel space. Therefore, it is desirable to modify the SOM to incorporate the relationship in the kernel space.

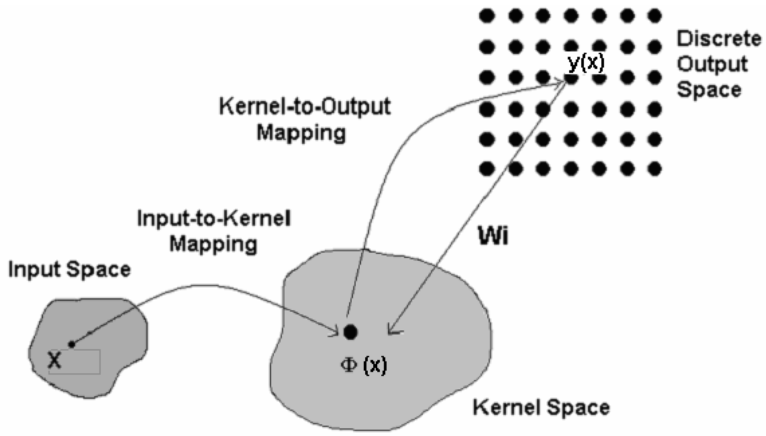

Figure 2. Original input space vs. transformed kernel space vs. output space.

As shown in Fig.2, the input sample **x** is transformed to **Φ** in the kernel space with proper kernel functions, and then mapped into **y** in the topology-preserving output space by the Self-Organizing Map. Therefore, the standard SOM learning rule in Eq.(2) [10]

$$\mathbf{w}_j[n+1] = \mathbf{w}_j[n] + \eta[n]h_{ji*}[n](\mathbf{x}[n] - \mathbf{w}_j[n]), \quad \mathbf{x} \in \Re^N \tag{2}$$

is modified to

$$\mathbf{w}_j[n+1] = \mathbf{w}_j[n] + \eta[n]h_{ji*}[n](\Phi(\mathbf{x}[n]) - \mathbf{w}_j[n]), \ \mathbf{w}_j \text{ and } \Phi \in \Re^M, \tag{3}$$

where $\Phi(\mathbf{x})$ is a kernel function to transform input space into the higher-dimensional kernel space with $N << M$. Here, $\mathbf{w}_j$ is the synaptic weight vector from the **Φ** to the j-th output neuron at the SOM, h_j is the neighborhood function from the winning node i^*, and η is the learning rate.

By representing the $\mathbf{w}_j$ as a linear superposition in the kernel space as

$$\mathbf{w}_j(n) = \sum_{k=1}^{L} a_{jk}^{(n)} \Phi(\mathbf{x}_k) \tag{4}$$

and substituting into Eq.(3), one obtains

$$\sum_{k=1}^{L} a_{jk}^{(n+1)} \Phi(\mathbf{x}_k) \bullet \Phi(\mathbf{x}_l) = (1 - \eta[n]h_{ji*}[n]) \sum_{k=1}^{L} a_{jk}^{(n)} \Phi(\mathbf{x}_k) \bullet \Phi(\mathbf{x}_l) + \eta[n]h_{ji*}[n]\Phi(\mathbf{x}) \bullet \Phi(\mathbf{x}_l). \tag{5}$$

Eq.(4) may be rewritten in a simple matrix form as

$$\mathbf{a}_j^{(n+1)}\mathbf{K} = (1 - \eta[n]h_{ji*}[n])\mathbf{a}_j^{(n)}\mathbf{K} + \eta[n]h_{ji*}[n]\mathbf{k}_{i*}, \tag{6}$$

where

$$\mathbf{K} \equiv \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_1) & \cdots & k(\mathbf{x}_L,\mathbf{x}_1) \\ k(\mathbf{x}_1,\mathbf{x}_2) & \ddots & & \vdots \\ \vdots & k(\mathbf{x}_j,\mathbf{x}_l) & \ddots & \vdots \\ k(\mathbf{x}_1,\mathbf{x}_L) & \cdots & \cdots & k(\mathbf{x}_L,\mathbf{x}_L) \end{bmatrix} \equiv \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \vdots \\ \mathbf{k}_L \end{bmatrix} = \begin{bmatrix} \mathbf{k}_1^T & \mathbf{k}_2^T & \cdots & \mathbf{k}_L^T \end{bmatrix} \tag{7}$$

with $k(\mathbf{x}_j,\mathbf{x}_l) \equiv \Phi(\mathbf{x}_j) \bullet \Phi(\mathbf{x}_l)$ and $\mathbf{a}_j^{(n)} \equiv [a_{j1}^{(n)} \quad a_{j2}^{(n)} \quad \cdots \quad a_{jL}^{(n)}]$. Then, using the winning node i^* with the current input **x**, the learning rule for the kernel-based SOM (K-SOM) is derived as

$$\mathrm{A}_{ji*}^{(n+1)} = (1 - \eta[n]h_{ji*}[n])\mathrm{A}_{ji*}^{(n)} + \eta[n]h_{ji*}[n] \ . \tag{8}$$

The winning node is determined as

$$i^*(\mathbf{x}) = \arg\min_j \| \Phi(\mathbf{x}) - \mathbf{w}_j[n] \|, \tag{9}$$

where

$$\| \Phi(\mathbf{x}) - \mathbf{w}_j(n) \|^2 = k(\mathbf{x},\mathbf{x}) + \mathrm{A}_j^{(n)} \mathbf{K}^T \{\mathrm{A}_j^{(n)}\}^T - 2\sum_{k=1}^{L} \alpha_{jk}^{(n)} k(\mathbf{x}_k, \mathbf{x}) \,. \tag{10}$$

Therefore, the winning node is selected for an input $\mathbf{x}_i$ as

$$i^*(\mathbf{x}_i) = \arg\min_j \{k(\mathbf{x}_i,\mathbf{x}_i) + \mathrm{A}_j^{(n)} \mathbf{K}^T \{\mathrm{A}_j^{(n)}\}^T - 2\mathrm{A}_j^{(n)} \mathbf{k}_i^T \} \,. \tag{11}$$

4. Results

To demonstrate the performance of the proposed SVM-BTA the recognition tasks are tested for 2 different databases, i.e., the CEDAR and TI46 databases, and the results are summarized in Table 1. The CEDAR is an image database for hand written digits from ‘0’ to ‘9’. The TI46 is a speech database for isolated words from ‘0’ to ‘9’ and ‘a’ to ‘z’. The tests are made with the SVM-light software at each node of the binary tree. [11]

The recognition rates of the SVM-BTA are compared with those of the MLP (Multi-layer Perceptron) and 2 popular multi-class SVM approaches. For the proposed SVM-BTA three different results are shown for the automatic grouping without overlap, automatic grouping with overlap, and the human grouping without overlap. Also, the results of grouping methods based on standard SOM and K-SOM are shown.

As expected, the “one-against-one” SVM approach shows much better recognition rates than MLP and the “one-against-the others” SVM approach. However, if the binary tree was designed from K-SOM with some overlaps, the proposed SVM-BTA shows similar or better performance than the “one-against-one” SVM approach. It is evident that both the kernel-based SOM and class overlaps are required for the best performance.

Table 1. Recognition Rates (%) of the Proposed SVM-BTA and Other Popular Classifiers

		MLP	SVM (1:1)	SVM (1:N-1)	SVM-BTA		
					No Overlap	Overlap	Human Grouping
CEDAR	SOM	93.6	94.4	90.0	80.8	84.8	
	K-SOM				90.6	94.6	94.7
TI46 (0-9)	SOM	90.0	97.7	96.6	96.4	96.5	
	K-SOM				96.6	97.6	98.0
TI46 (0-9,a-z)	SOM		89.6	72.5	66.2	67.3	
	K-SOM				83.4	90.5	

5. Conclusions

The Support Vector Machines with Binary Tree Architecture (SVM-BTA) are designed to demonstrate superior multi-class classification performance than other classifiers including the much-demanding “one-against-one” SVM approach. Also, due to the tree architecture, the SVM-BTA requires much less computation for decision making. However, it is important to convert the multi-class problems into proper binary trees, and the Kernel-based SOM may be utilized for this conversion.

The conversion is automated by maximizing the scattering measure at the kernel space, but currently requires exhaustive search for all possible combinations. Although this conversion is done only once at the classifier design stage, an efficient grouping algorithm may be better to have and is subject to the future researches.

Acknowledgment: This research was supported as a Brain Neuroinformatics Research Program by Korean Ministry of Science and Technology.

References

[1] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[2] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 1995.

[3] C.J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery,* Vol2, pp. 121-167, 1998.

[4] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning, Vol*. 20, No. 3, pp. 273-297, 1995.

[5] K.-R. Muller, S. Mika, G. Ratisch, K. Tsuda, and B. Scholkopf, “An Introduction to Kernel-Based Learning Algorithms,” *IEEE Trans. Neural Networks*, Vol. 12, No. 2, 2001.

[6] B. Scholkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A.J. Smola, “Input Space vs. Feature Space in Kernel-Based Methods,” *IEEE Trans. Neural Networks*, Vol. 10, No. 5, pp. 1000-1017, 1999.

[7] F. Schwenker, “Hierarchical Support Vector Machines for Multi-Class Pattern Recognition,” *Fourth International conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, Brighton, UK*, 2000.

[8] J. Weston and C. Watkins, Multi-class support vector machines, *Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, Egham,* UK, 1998.

[9] Y. Lee, Y. Lin, and G. Wahba, “Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data,” *Journal of Atmospheric and Oceanic Technology*, 2003.

[10] S. Haykin, *Neural networks: A Comprehensive Foundation*, Prentice Hall, 2nd Ed., 1999.

[11] T. Joachims, Making Large-Scale SVM Learning Practical, *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge, USA, 1998.

Sungmoon Jeong had graduated with MS from the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, and now is working for the Electronics and Telecommunications Research Institute, Daejeon, Korea.

Sang-Hoon Oh had graduated with MS from Pusan National University and Ph.D. from the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology. He is now an Assistant Professor at the Department of Electrical Engineering, University of Mokwon, Daejeon, Korea.

Soo-Young Lee had graduated with BS from Seoul National University in 1975, MS from Korea Advanced Institute of Science and Technology in 1977, and Ph.D. in Electrophysics from Polytechnic Institute of New York in 1984. He joined the Korea Advanced Institute of Science and Technology as an Assistant Professor in 1986, and now is a Professor at the Department of BioSystems and Director of the Brain Science Research Center. His current research interests include modeling, applications, and chip implementations of human auditory pathways for robust speech feature extraction, binaural processing, and top-down attention.