

Error Correcting Output Codes vs. Fuzzy Support Vector Machines

Tomonori Kikuchi

Electrical and Electronics Engineering, Kobe University
Kobe, Japan

Shigeo Abe

Graduate School of Science and Technology, Kobe University
Kobe, Japan
abe@eedept.kobe-u.ac.jp

Abstract

Error correcting output codes (ECOC) have been proposed to enhance generalization ability of classifiers. If, instead of discrete error functions, continuous error functions are used, unclassifiable regions of multiclass support vector machines are resolved. In this paper, we discuss minimum operations as well as average operations for error functions of support vector machines and show the equivalence of ECOC support vector machines and fuzzy support vector machines for one-against-all formulation. Then we show by computer simulations that ECOC support vector machines are not always superior to one-against-all fuzzy support vector machines.

1. Introduction

Since support vector machines are formulated for two-class classification problems [1], an extension to multi-class problems is not unique. Original formulation by Vapnik [1] is one-against-all classification, in which one class is separated from the remaining classes. By this formulation, however, unclassifiable regions exist. Instead of discrete decision functions, Vapnik [2, p. 438] proposed to use continuous decision functions. Namely, we classify a datum into the class with the maximum value of the decision functions. Inoue and Abe [3] proposed fuzzy support vector machines, in which membership functions are defined using the decision functions. Abe [4] showed that support vector machines with continuous decision functions and fuzzy support vector machines are equivalent.

Dietterich and Bakiri proposed error correcting output codes (ECOC) to enhance generalization ability of classifiers [5] borrowing the idea of error correcting codes used for correcting bit errors in transmission channels. One-against-all formulation is a special case of error correcting codes with no error correcting capability, and by introducing “don’t” care bits, also is pairwise formulation

[6]. Using the continuous Hamming distance for support vector machines, instead of the Hamming distance, unclassifiable regions are resolved.

In this paper, we discuss ECOC support vector machines in comparison to fuzzy support vector machines. First we define the distance between codes by the maximum of continuous error functions as well as the continuous Hamming distance, which is the sum of continuous error functions. Next we show that these definitions are equivalent to the fuzzy support vector machines with minimum and average operators, respectively. Then we compare recognition performance of ECOC support vector machines with that of one-against-all fuzzy support vector machines.

In Section 2, we explain two-class support vector machines, and in Section 3 we discuss ECOC support vector machines. In Section 4, we compare recognition performance of ECOC support vector machines with average and minimum operators with one-against-all fuzzy support vector machines.

2. Two-class Support Vector Machines

Let m -dimensional inputs \mathbf{x}_i ($i = 1, \dots, M$) belong to Class 1 or 2 and the associated labels be $y_i = 1$ for Class 1 and -1 for Class 2. Let the decision function be

$$D(\mathbf{x}) = \mathbf{w}^t \mathbf{x} + b, \quad (1)$$

where \mathbf{w} is an m -dimensional vector, b is a scalar, and

$$y_i D(\mathbf{x}_i) \geq 1 - \xi_i \quad \text{for } i = 1, \dots, M. \quad (2)$$

Here ξ_i are nonnegative slack variables.

The distance between the separating hyperplane $D(\mathbf{x}) = 0$ and the training datum, with $\xi_i = 0$, nearest to the hyperplane is called margin. The hyperplane $D(\mathbf{x}) = 0$ with the maximum margin is called optimal separating hyperplane.

To determine the optimal separating hyperplane, we minimize

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^M \xi_i \quad (3)$$

subject to the constraints:

$$y_i (\mathbf{w}^t \mathbf{x}_i + b) \geq 1 - \xi_i \quad \text{for } i = 1, \dots, M, \quad (4)$$

where C is the margin parameter that determines the tradeoff between the maximization of the margin and minimization of the classification error. The data that satisfy the equality in (4) are called support vectors.

To enhance separability, the input space is mapped into the high-dimensional dot-product space called feature space. Let the mapping function be $\mathbf{g}(\mathbf{x})$. If the dot product in the feature space is expressed by $H(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\mathbf{x})^t \mathbf{g}(\mathbf{x}')$, $H(\mathbf{x}, \mathbf{x}')$ is called kernel function, and we do not need to explicitly treat the feature space. The kernel functions used in this study are as follows:

1. Polynomial kernels

$$H(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^t \mathbf{x}' + 1)^d, \quad (5)$$

where d is an integer.

2. RBF kernels

$$H(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2), \quad (6)$$

where γ is a positive parameter for slope control.

3. Error-correcting Output Codes

Error correcting codes, which detect and correct errors in data transmission channels, are used to improve generalization ability in pattern classification. For support vector machines, in addition to generalization improvement they can be used to resolve unclassifiable regions. In this section, first we discuss how error-correcting codes can be used for pattern classification. Next, by introducing “don’t care” output, we discuss a unified scheme for output coding that includes one-against-all and pairwise formulations [6]. Then we show the equivalence of the error correcting codes with the membership functions.

3.1. Output Coding by Error-correcting Codes

Dietterich and Bakiri [5] proposed to use error-correcting output codes for multiclass problems. Let g_{ij} be the target value of the j th decision function $D_j(\mathbf{x})$ for class i :

$$g_{ij} = \begin{cases} 1 & \text{if } D_j(\mathbf{x}) > 0 \text{ for class } i, \\ -1 & \text{otherwise.} \end{cases} \quad (7)$$

The j th column vector $\mathbf{g}_j = (g_{1j}, \dots, g_{nj})^t$ is the target vector for the j th decision function, where n is the number of classes. If all the elements of a column are 1 or

Table 1. Error-correcting codes for three classes

Class	\mathbf{g}_1	\mathbf{g}_2	\mathbf{g}_3
1	1	-1	-1
2	-1	1	-1
3	-1	-1	1

Table 2. Extended error-correcting codes for pairwise classification with three classes

Class	\mathbf{g}_1	\mathbf{g}_2	\mathbf{g}_3
1	1	0	-1
2	-1	1	0
3	0	-1	1

-1, classification is not performed by this decision function and two column vectors with $\mathbf{g}_i = -\mathbf{g}_j$ result in the same decision function. Thus the maximum number of distinct decision functions is $2^{n-1} - 1$.

The i th row vector (g_{i1}, \dots, g_{ik}) corresponds to a code word for class i , where k is the number of decision functions. In error-correcting codes, if the minimum Hamming distance between pairs of code words is h , the code can correct at least $\lfloor (h-1)/2 \rfloor$ bit errors. For 3-class problems, there are three decision functions in maximum as shown in Table 1, which is equivalent to one-against-all formulation and there is no error-correcting function. Thus ECOC is considered to be a variant of one-against-all classification.

3.2. Unified Scheme for Output Coding

Introducing “don’t care” outputs, Allwein, Schapire, and Y. Singer [6] unified output codes that include one-against-all, pairwise, and ECOC schemes. Denoting a “don’t care” output by 0, pairwise classification [7] for three classes can be shown as in Table 2.

To calculate the distance of \mathbf{x} from the j th decision function for class i , we define the error $\varepsilon_{ij}(\mathbf{x})$ by

$$\varepsilon_{ij}(\mathbf{x}) = \begin{cases} 0 & \text{for } g_{ij} = 0, \\ \max(1 - g_{ij} D_j(\mathbf{x}), 0) & \text{otherwise.} \end{cases} \quad (8)$$

If $g_{ij} = 0$, we need to skip this case. Thus, $\varepsilon_{ij}(\mathbf{x}) = 0$. If $g_{ij} D_j(\mathbf{x}) \geq 1$, \mathbf{x} is on the correct side of the j th decision function with more than or equal to the maximum margin. Thus, $\varepsilon_{ij}(\mathbf{x}) = 0$. If $g_{ij} D_j(\mathbf{x}) < 1$, \mathbf{x} is on the wrong side

or even if it is on the correct side, the margin is smaller than the maximum margin. We evaluate this disparity by $1 - g_{ij}D_i(\mathbf{x})$.

Then the distance of \mathbf{x} from class i is given by

$$d_i(\mathbf{x}) = \sum_{j=1}^k \varepsilon_{ij}(\mathbf{x}). \quad (9)$$

Using (9), \mathbf{x} is classified into

$$\arg \min_{i=1, \dots, n} d_i(\mathbf{x}). \quad (10)$$

Instead of (8), if we use the discrete function:

$$\varepsilon_{ij}(\mathbf{x}) = \begin{cases} 0 & \text{for } g_{ij} = 0, \\ 0 & \text{for } g_{ij} = \pm 1, g_{ij}D_i(\mathbf{x}) \geq 1, \\ 1 & \text{otherwise,} \end{cases} \quad (11)$$

(9) gives the Hamming distance. But by this formulation unclassifiable regions occur.

3.3. Equivalence of ECOC with Membership Functions

Here, we discuss the relationship between ECOC and membership functions. For $g_{ij} = \pm 1$, the error $\varepsilon_{ij}(\mathbf{x})$ is expressed by the one-dimensional membership functions $m_{ij}(\mathbf{x})$:

$$\begin{aligned} m_{ij}(\mathbf{x}) &= \min(g_{ij}D_j(\mathbf{x}), 1) \\ &= 1 - \varepsilon_{ij}(\mathbf{x}). \end{aligned} \quad (12)$$

Thus, if we define the membership function for class i by

$$m_i(\mathbf{x}) = \frac{1}{\sum_{j=1}^k |g_{ij}|} \sum_{\substack{j=1 \\ g_{ij} \neq 0}}^k m_{ij}(\mathbf{x}) \quad (13)$$

and classify \mathbf{x} into

$$\arg \max_{i=1, \dots, n} m_i(\mathbf{x}), \quad (14)$$

we obtain the same recognition result as that by (10). This is equivalent to a fuzzy support vector machine with the average operator.

Similarly, instead of (9), if we use

$$d_i(\mathbf{x}) = \max_{j=1, \dots, n} \varepsilon_{ij}(\mathbf{x}), \quad (15)$$

the resulting classifier is equivalent to a fuzzy support vector machine with minimum operators.

4. Performance Evaluation

We evaluated recognition performance of ECOC support vector machines with one-against-all support vector machines using the blood cell data and hiragana data

listed in Table 3. The blood cell classification involves classifying optically screened white blood cells into 12 classes [8]. This is a very difficult problem; class boundaries for some classes are ambiguous because the classes are defined according to the growth stages of white blood cells. Hiragana data are gathered from Japanese license plates. The original gray-scale images of hiragana characters were transformed into 5×10 -pixel with the gray-scale range being from 0 to 255. Then by performing gray-scale shift, position shift, and random noise addition to the images, the training and test data were generated [9]. Hiragana data are relatively easy to be classified.

Table 3. Benchmark data specification

Data	Inputs	Classes	Train.	Test
Blood cell	13	12	3097	3100
Hiragana	50	39	4610	4610

As error correcting codes we used the BCH (Bose-Chaudhuri-Hochquenghem) codes, which belong to one type of cyclic codes. We used four BCH codes with 15, 31, 63, and 127 word lengths, properly setting the minimum Hamming distances. For each word length we trained 10 ECOC support vector machines with $C = 5000$ changing code words.

Table 4 shows the results for the blood cell data with polynomial kernels with degree 3. In the ‘‘Code’’ column, e.g., (15, 7) means that the word length is 15 bits and the minimum Hamming distance is 7. The ‘‘Hamming,’’ ‘‘Average,’’ and ‘‘Minimum’’ columns list the average recognition rate of the test and the training data (in the brackets) using the Hamming distance, the average operator, and the minimum operator, respectively. The boldfaced numeral shows the maximum recognition rate among different codes.

From the table, the recognition rates of both training and test data improved as the word length was increased and they reached the maximum at the word length of 63. But since by the Hamming distance unclassifiable regions exist, the recognition rates are lower than by average and minimum operators. By the average and minimum operators, however, the one-against-all support vector machines showed the best recognition rates. This may be caused by the lower recognition rates of the training data for the ECOC support vector machines than by the one-against-all support vector machines.

Thus, to improve the recognition rate of the test data, we used the RBF kernels. Table 5 shows the results for the RBF kernels with $\gamma = 1$. The ECOC support vector machines showed better recognition performance than the one-against-all support vector machines. In addition, the average operator showed better recognition performance than the minimum operator.

Table 6 shows the results of hiragana data for the polynomial kernels with degree 3. The ECOC support vector

Table 4. Recognition rates (%) of blood cell data with polynomial kernels ($d = 3$)

Code	Hamming	Average	Minimum
1-all	87.13 (92.41)	92.84 (96.09)	92.84 (96.09)
(15,7)	90.17 (93.34)	91.56 (94.45)	91.19 (93.95)
(31,11)	90.86 (93.60)	91.90 (94.59)	91.80 (94.16)
(63,31)	91.82 (94.64)	92.20 (94.98)	92.23 (94.32)
(127,63)	91.80 (94.58)	92.01 (94.82)	91.93 (96.09)

Table 5. Recognition rates (%) of blood cell data with RBF kernels ($\gamma = 1$)

Code	Hamming	Average	Minimum
1-all	86.68 (98.58)	92.94 (99.29)	92.94 (99.29)
(15,7)	92.43 (98.27)	93.47 (98.49)	93.07 (98.18)
(31,11)	92.88 (98.36)	93.85 (98.59)	93.53 (98.13)
(63,27)	93.68 (98.64)	94.05 (98.68)	93.75 (98.37)
(127,55)	93.68 (98.60)	93.96 (98.61)	93.63 (97.94)

machine with the average operator and the word length of 127 showed the best recognition performance. But some ECOC support vector machines showed lower recognition performance than the one-against-all support vector machine. Thus the performance of ECOC support vector machines was not stable.

Unlike blood cell data, since the recognition rates of the training data are near 100% using polynomial kernels, the improvement using RBF kernels was not recognized. Thus, we do not include the results here.

Table 6. Recognition rates (%) of hiragana data with polynomial kernels ($d = 3$)

Code	Hamming	Average	Minimum
1-all	99.28 (100)	99.28 (100)	99.28 (100)
(15,7)	95.50 (99.96)	97.63 (99.85)	96.93 (99.97)
(31,11)	98.38 (99.99)	99.01 (100)	98.56 (99.97)
(63,31)	99.01 (100)	99.30 (100)	99.17 (99.97)
(127,63)	99.31 (100)	99.46 (100)	99.26 (99.97)

4.1. Discussions

For the blood cell data, ECOC support vector machines showed better performance than fuzzy support vector machines when RBF kernels were used, but for the hiragana data, improvement was not significant if any. This is because for the hiragana data fuzzy support vector machines could achieve relatively good performance and little was left for improvement.

For the blood cell data with polynomial kernels and for the hiragana data, ECOC support vector machines did not perform better than one-against-all support vector machines. And for the hiragana data, the recognition performance against the length of code word was unstable. Thus, to obtain good recognition performance, we need to optimize the structure of ECOC support vector machines.

5. Conclusions

In this paper, we discussed minimum operators as well as average operators for the error functions of ECOC support vector machines and show the equivalence of ECOC support vector machines and fuzzy support vector machines for one-against-all formulation. By computer simulations we showed that ECOC support vector machines are not always superior to one-against-all fuzzy support vector machines.

References

- [1] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, London, UK, 1995.
- [2] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, NY, 1998.
- [3] T. Inoue and S. Abe. Fuzzy support vector machines for pattern classification. In *Proceedings of International Joint Conference on Neural Networks (IJCNN '01)*, volume 2, pages 1449–1454, July 2001.
- [4] S. Abe. Analysis of multiclass support vector machines. In *Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2003)*, pages 385–396, Vienna, Austria, February 2003.
- [5] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.
- [6] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

- [7] U. H.-G. Kreßel. Pairwise classification and support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 255–268. The MIT Press, Cambridge, MA, 1999.
- [8] A. Hashizume, J. Motoike, and R. Yabe. Fully automated blood cell differential system and its application. In *Proceedings of the IUPAC Third International Congress on Automation and New Technology in the Clinical Laboratory*, pages 297–302, Kobe, Japan, September 1988.
- [9] M.-S. Lan, H. Takenaga, and S. Abe. Character recognition using fuzzy rules extracted from data. In *Proceedings of the Third IEEE International Conference on Fuzzy Systems*, volume 1, pages 415–420, Orlando, FL, June 1994.