Using Hadamard ECOC in multi-class problems based on SVM

Yin An-rong, Xie Xiang, Kuang Jing-ming

Dept. of Electronics Engineering
Beijing Institute of Technology, Beijing, China
{anrong, xiexiang, jmkuang}@bit.edu.cn

10.21437/Interspeech.2005-672

Abstract

In this paper, we propose to apply Hadamard Error-Correcting Output Code (Hadamard ECOC) to extend binary classifier for multi-class classification problems. Hadamard ECOC is easy to construct and is suitable for any number of classes. We combine it with binary support vector machine (SVM) to solve the multi-class problem of speaker identification, which takes advantage of error correcting ability of Hadamard ECOC and powerful classification ability of SVM. Compared to the traditional "1-against-rest" method, the experiment result shows that Hadamard ECOC approach has much better and more stable performance for the multi-class problem and is robust on different rules mapping rules between ECOCs and classes.

1. Introduction

How to process multi-class problem has been one of the research focuses in pattern recognition for years. Exiting approaches include direct application of multi-class algorithms such as the decision-tree algorithm, application of binary concept learning algorithms to learn individual binary classifier for each of the classes [1], application of binary concept learning algorithms with distribution output representation [2] and so on. One of the most efficient approaches is the application of binary concept learning algorithms with error-correcting output codes (ECOC) which is proposed by Dietterich and Bakiri [3]. They tested this method on variant multi-class tasks combined with decision-tree and artificial neural networks learning algorithms. The experiment results demonstrated that ECOC provided a general-purpose method for improving the performance of learning algorithms on multi-class problems and was superior to the other approaches.

However, they didn't give a single method to construct ECOC suitable for any number of classes, *k*, and one has to construct different ECOC via different algorithms according to the different values of *k*, that is, the algorithms have to be changed as *k* changes. Moreover, the algorithms proposed above are complex.

This paper proposes a kind of general ECOC, Hadamard ECOC, which is not only much more convenient to generate but also is suitable for any number of classes. The experiment result shows that it can extend binary classifier to multi-class problems efficiently.

The binary classifier applied in this paper is support vector machine (SVM) [4], which is now very popular in pattern recognition because of its powerful classification ability. The other approaches to extend it to multi-class problems include "1-against-1", "1-against-rest" and so on [5].

The remainder of this paper is organized as follows: Section 2 introduces SVM; Sec. 3 and Sec. 4 describes the basic principle of ECOC and the construction of Hadamard ECOC respectively; Sec. 5 is dedicated to the combination of ECOC and the binary classifier SVM; The experimental task, data and results are presented in Sec. 6 with the conclusion in Sec. 7.

2. Support vector machine

SVM is proposed by V. Vapnik and developed from the statistical learning theory [4], with the aim of predicting the unknown output accurately by evaluating the relationship between the known input and output according to the given training samples.

SVM is originally designed for binary classification, and its target is to find an optimal separating hyperplane *w*, which can minimize the number of errors made on the training set while simultaneously maximizing the margin between the individual classes. When a new example, *x*, is introduced, the decision result about it will be gained by the following decision function:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} y_i \partial_i K(x, x_i) + b) \qquad (1)$$

Where ∂_i is the non-zero Lagrange Coefficients, x_i is the support vectors, y_i is the corresponding class label of x_i, $y_i \in \{0,1\}$, and $K(.,.)$ is the kernel function. In the linearly non-separable case, SVM will map the input data to a high dimensional feature space to find a linear hyperplane via kernel function, which is also the key feature of SVM.

There are many kinds of kernel functions, such as polynomial kernel, Fisher kernel and sigmoid kernel, and radial base function (RBF) is used in this paper:

$$K(x, x_i) = \exp\{-r \| x - x_i \|^2\} \qquad (2)$$

As discussed above, there are many implementations for SVM multi-class classification such as "1-aginst-rest" and "1-against-1" [5]. For "1-against-rest" approach, we build *k* SVM models where *k* is the number of classes. The *i*th SVM is trained with all of the examples in the *i*th class with label "1" and all other examples with label "0". In this paper, we compare the performance between "1-against-rest" approach and Hadamard ECOC.

3. Error-correcting output code

ECOC was introduced by Bose and Ray-Chaudhuri in 1960 [6]. The idea of employing ECOC to machine learning was proposed by Duda，Machanik and Singleton in 1963 [7]. Dietterich and Bakiri proposed ECOC to extend binary

classifier to solve multi-class problem with the number of classes k in 1995 [3]: Each class is assigned a unique binary string W_i (codeword) of length L, so we get a codebook with k rows and L columns. During training, the training set of classes will be partitioned into complementary (two) subsets according to the labels of each column and each such partition defines a binary problem which is used to train a binary classifier for each column. To classify a new example, x, the L binary classifiers are evaluated on x to obtain a binary sequence $B = \{b_1, b_2, ..., b_L\}$. Then the distances of this sequence to each of the k codewords are computed. The class corresponding to the nearest codeword, according to Hamming distance, is the decision result:

$$\widehat{W} = \arg\min d(B, W_i) = \sum_{j=1}^{L} | b_j - W_{i,j} | \qquad (3)$$

Any approach to solve classification problem consists of the following procedures: collecting training samples, extracting feature and implementing the learning algorithm. Because of errors introduced by the finite training samples, poor choice of feature and flaws in the learning process, the class information is corrupted, which maybe lead to classification errors. By encoding the classes by ECOC, the system may be able to recover from the errors. If the minimum Hamming distance between any pair of code words of ECOC is d, then the code can correct $\lfloor (d-1)/2 \rfloor$ single bit errors. That is, if the classifiers make only $\lfloor (d-1)/2 \rfloor$ errors, the nearest codeword will still be the correct codeword and the decision result will still be right.

When applying ECOC to extend binary classifiers to multi-class problems, we should choose ECOC which satisfies the following requirements [3]:

1， Row separation. Each codeword should be well-separated in Hamming distance from each of the other codewords, because the power of a code to correct errors is directly related to the row separation, as discussed above.

2， Column separation. In order to reduce the correlation between any pair of binary classifiers, each column should be well-separated in Hamming distance from each of the other columns and from each of the complement of the other columns. If columns are similar or identical, then when a learning algorithm such as SVM is applied to learn the classifiers, it will make similar (correlated) mistakes, which may lead to many simultaneous errors which can not be corrected during test. If columns are complementary, the binary classifiers trained according to these columns may be identical to each other. This is because some algorithms such as SVM treat a class and its complement symmetrically.

Based on the above two requirements, Dietterich and Bakiri employed four methods for designing ECOC according to the number of classes, k: Exhaustive codes ($3 \leqslant k \leqslant 7$), Column selection from exhaustive codes ($8 \leqslant k \leqslant 11$), Randomized hill climbing algorithm ($k \geqslant 11$) and BCH codes ($k \geqslant 11$) [3]. But there are three problems:

1， They didn't propose a single method suitable for all values of k, which is not convenient for code design.

2， It is very complex to apply these algorithms. Take BCH codes for example, it is difficult to construct such codes. And even we have got the code matrix, we have to experiment with certain algorithm for code shortening and column selection because the number of rows is always a power of two and there are complementary columns.

3， The codeword is very long, which means a large number of binary classifiers have to be constructed. Take Exhaustive codes for example, the length of the code, L, is 2^{k-1}-1 where k is the number of classes, so the code length is as long as 63 just for 7-class problem, which leads to the improvement of the performance but at too great cost of both storage and computation.

Hadamard ECOC is proposed to solve the above three problems in this paper.

4. Hadamard ECOC

Hadamard ECOC is derived from Hadamard matrix. The Hadamard matrix of second order is given by:

$$H_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad (4)$$

If N is the power of two, symmetrical Hadamard matrix with rank N can be defined recursively as follows [8]:

$$H_N = \begin{pmatrix} H_{N/2} & H_{N/2} \\ H_{N/2} & -H_{N/2} \end{pmatrix} \qquad (5)$$

Where $-H_{N/2}$ is the complement of $H_{N/2}$. So the rows and columns of Hadamard matrix are orthogonal with each other (In order to be consistent with the code words discussed above, we used "0" instead of "-1" for the elements of Hadamard matrix). Moreover, the distance between any pair of rows or columns is $N/2$ in Hadamard matrix with rank N. Because of those characters introduced above, Hadamard matrix is popular in variant fields such as encoding theory, communication and DSP. In addition, we can notice that the separation between rows and columns of Hadamard satisfies the both two requirements about ECOC described in Sec. 3. However, Hadamard matrix can't be applied to multi-class problem directly because of the following two reasons:

1， The first column of Hadamard matrix is all zero, which means that there is no binary classifier according to this column. Hence it is useless for the ECOC matrix and should be discarded.

2， The number of classes is restricted. Hadamard matrix is a $2^j \times 2^j$ square matrix, so it is only able to encode 2^j classes. We have to select suitable ECOC from the matrix when we want to encode any classes.

After some processing, Hadamard ECOC for k-class can be obtained by the following steps:

1， Create Hadamard matrix with rank N in Equation (5). Let $N=2^j$ if $2^{j-1} < k \leq 2^j$.

2， Get a $2^j \times 2^j$-1 square matrix by deleting the first column of Hadamard matrix.

3， Derive the required $k \times 2^j$-1 ECOC array from the matrix in step2 by selecting the first k rows directly.

One can get Hadamard ECOC for any number of classes via the method described above. This method has following three advantages:

1， Simple and rapid. It is easy to create Hadamard matrix and convenient to derive the ECOC codebook from Hadamard matrix.

2， The code length of resulting Hadamard ECOC is 2^j-1.

Table 1: A 15-bits Hadamard error-correcting output code for a ten-class problem

class	codewords														
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
2	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
3	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
4	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
5	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0
6	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
7	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1
8	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
9	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0

The hamming distance between any pair of rows is 2^{j-1} and there are no identical or complemental columns, which guarantees the row separation and the column separation.

3， Compared with ECOC introduced in Sec. 3, the codeword of Hadamard ECOC is much shorter (the code length L is always between k-1 and $2k$-1) and thus the number of binary classifiers is much fewer, which efficiently reduces the storage and improves the running efficiency of the multi-class classifier.

Table 1 shows a 15-bit Hadamard ECOC for a 10-class task and the minimum hamming distance between any pair of code words is 8.

5. Combination of ECOC and SVM

When combined with support vector machine, each column of ECOC is corresponding to a SVM, that is, a SVM is constructed for each column via the training data of classes labeled "0" against the training data of classes labeled "1" during training.

To classify a new example $X = \{x_1, x_2, ..., x_T\}$, which is a feature parameter sequence extracted from the test data, e.g. a speech segment, each of the SVMs is evaluated on X and outputs two values about each frame x_i: the decision result "0" or "1" of the decision function (i.e. the label of this frame, see Eq. (1)) and the value of the decision function before the decision function takes "sgn". How to get the final decision result according to the output of each SVM? There are two approaches:

1， Let the output of each SVM be "0" or "1" according to some rule such as the ratio of the number of the frames labeled "0" or "1" to the number of total frames in the test example. Then compute the distance of the outputs of all SVMs, which is often a binary string, to each of the k codewords. The class corresponding to the nearest codeword, according to Hamming distance, is the decision result (see Eq. (3)). This is a hard decision.

2， Let the output of each SVM present the probability that the current test example belongs to class "0" or "1", which can be the ratio of the number of the frames labeled "0" or "1" to the number of total frames in the test example. Then compute the distances of outputs of all SVMs, $P = \{p_1, p_2, ..., p_L\}$, to the ith codeword by[3]:

$$d(P, W_i) = \sum_{j=1}^{L} | p_j - W_{i,j} | \quad (6)$$

The class corresponding to the farthest codeword is the decision result if p_i present the probability that the test example belongs to class "0"; the class corresponding to the nearest codeword is the decision result if p_i present the probability that the test example belongs to class "1". This is a soft decision.

In this paper, we adopt the second method but use the sum of the absolute value $g(x)$ of the decision function before it takes "sgn" instead of the number of frames to present the probability:

$$g(x) = | \sum_{i=1}^{n} y_i \partial_i K(x, x_i) + b | \quad (7)$$

Which indicates the distance of the new example to the classifying plane and measures the confidence of this decision.

Assumed that the assembly of the frames labeled "0" ("1") by the decision function of the lth SVM is $S \subseteq \{1, 2, ..., T\}$, then the probability that the test example belongs to class "0" ("1") is given by

$$p_l = \sum_S g(x_i) \Big/ \sum_{i=1}^{T} g(x_i) \quad (8)$$

Because the decision result ("0" or "1") given by the decision function in Eq. (3) doesn't take account of the magnitude of $g(x)$ (i.e. the confidence of this decision), Eq. (8) present the probability that test example belongs to class "0" ("1") more essentially than the ratio of frames' number.

6. Experiments

6.1. Experimental setup

The speech database used in this paper is collected over the telephone channel in relative quiet environment by Modern Communication Lab of BIT, with 8k sample rate and 16 bits per sample. It contains 15 speakers (8 men and 7 women) with 40 utterances for each speaker. Each utterance is a short sentence with the length of about 3s.

The feature vectors used in this paper comprise of 12 Mel-Frequency Cepstral Coefficient (MFCC) derived from 25 filter banks. Each feature vector is extracted at 10ms intervals using a 20ms hamming window.

6.2. Results

We test the proposed approach on a wide range of multi-class tasks of text-independent speaker identification from 7-class to 15-class (i.e. 7 speakers to 15 speakers in a close set). To take full advantage of the database, cross-validation method is used, i.e. 20 utterances are used as the training set with the rest 20 utterance as the test set and then in reverse for each speaker, so the training speech is about 1 minute and the test speech is about 3s for each trial. The number of trials is $M=k\times 2\times 20$ for k-class problem. The number of the male speakers is identical to that of the female speakers when the number of total speaker is even and the number of the male speakers is one more than that of the female speakers when the number of total speaker is odd.

We apply both "1-against-rest" and Hadamard ECOC approaches to solve the same multi-class problems with the binary classifier SVM. The comparison of test results between the "1-against-rest" method and Hadamard ECOC is shown in Figure 1.

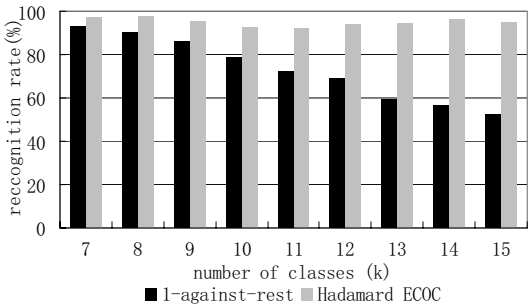

Figure 1: Recognition rates of "1-against-rest" method and Hadamard ECOC

Figure 1 shows:

1， Hadamard ECOC is always superior to "1-againt-rest" when they are applied to solve the same multi-class problem. It is more obvious as the number of classes becomes larger, e.g. Hadamard ECOC approach improves the recognition rate by about 79%, from 52.83% to 94.83% when the number of classes is 15. In fact, "1-against-rest" can be regarded as a special ECOC whose code book is an identity matrix of which the minimum hamming distance between any pair of rows is 2, so it can't correct any errors and is inferior.

2， The performance of "1-against-rest" reduces rapidly as the number of classes increases. This is because discrimination among "1-against-rest" models reduces as the number of speakers increases. As a contrast, the performance of Hadamard ECOC is much more stable.

In the results reported above, the codewords in the Hadamard ECOC have been arbitrarily assigned to the classes (speakers). Is Hadamard ECOC sensitive to the mapping rules between classes and codewords? We conduct a series of experiments on 7-class problem to determine the question. Table 2 shows the results of five random assignments of codewords to classes. We can observe that there is no statistically significant variation in the performance of the different random assignments.

From the experiment results above, we can notice that Hadamard ECOC approach is much superior to "1-against-rest" and is more substantial to any number of classes. Moreover, it is robust with respect to the different mapping rules between codewords and classes.

Table 2: The results of five random assignments of codewords to classes

1-against-rest (%)	5 assignments of 7-bitsHadamard ECOC (%)				
	a	b	c	d	e
93.21	97.14	96.43	97.86	97.14	96.43

However, there is some additional cost to employ ECOC. SVMs using ECOC are generally larger and more complex than SVMs constructed using "1-againt-rest" approach when k is not the power of 2. So ECOC method is not appropriate in the domains where training must be rapid.

7. Conclusions

In this paper, we have proposed a new kind of ECOC, Hadamard ECOC, which is convenient to construct and suitable to any number of classes for multi-class problems. We have tested it on a wide range of multi-class tasks of text-independent speaker identification with binary SVM. Moreover, we have experimentally compared this approach with the classical "1-against-rest" approach. The results clearly show that the Hadamard ECOC can extend the binary classifier SVM to multi-class problem efficiently and is robust with respect to the random assignment of codewords to classes. How to improve the efficiency of ECOC multi-classifier to make it more suitable for practical use is our future research.

8. Acknowledgememts

This work was supported in part by the National Nature Science Foundation of P.R.China under Grant NSFC 60372089.

9. References

[1] Nilsson, N. J., *Learning Machines*. McGraw-Hill, New York. 1965.

[2] Sejnowski, T. J., Rosenberg, C. R., "Parallel networks that learn to pronounce English text", Journal of Complex System, 1(1): 145-168, 1987.

[3] T. G. Dietterich and G. Bakiri. "Solving multiclass learning problems via error-correcting output codes", Journal of Artificial Intelligence Research, 2: 263 – 286, 1995

[4] V. N. Vapnik. *The nature of statistical learning theory*[M]. Springer. 1995.

[5] Chih-Wei Hsu and Chin-Jen Lin "A comparison of methods for multiclass support vector machine"，IEEE Transactions on Neural Network, vol.13, No2：415-425, 2002.

[6] Bose, R. C., Ray-Chaudhuri, D. K. "On a class of error-correcting binary group codes", Information and Control, 3: 68-79, 1960.

[7] Duda, R. O., Machanik, J. w.,Singleton, R. C. "Function modeling experiments", Tech. rep. 3605, Stanford Research Institute, 1963.

[8] W. K. Pratt, J. Kane and H. C. Andrews, "Hadamard transform image encoding", Proc. IEEE, Vol. 57, pp. 58-68, Jan 1969.