Optimisation of Multiclass Supervised Classification Based on Using Output Codes with Error-Correcting[1,2]

V. V. Ryazanov

Department of Computer Science, Moscow Institute of Physics and Technology, Moscow, Russia
e-mail: vasyarv@mail.ru

Abstract—An approach of solving the problem of multiclass supervised classification, based on using error-correcting codes is considered. The main problem here is the creation of binary code matrix, which provides high classification accuracy. Binary classifiers must be distinct and accurate. In this issue, there are many questions. What should be the elements of the matrix, how many elements provide the best accuracy and how to find them? In this paper an approach to solve some optimization problems for the construction of the binary code matrix is considered. The problem of finding the best binary classifiers (columns of matrix) is formulated as a discrete optimization problem. For some partial precedent classification approach, there is a calculation of the effective values of optimising function. Prospects of this approach are confirmed by a series of experiments on various practical tasks.

Keywords: classification, data mining, supervised learning, multiclass, codeword.

DOI: 10.1134/S1054661816020176

1. INTRODUCTION

In this paper the problem of multiclass classification is considered. A training set containing representatives of a finite number of classes is assumed to be given. The task is to create a classification algorithm, which classifies a new arbitrary object to one of the classes. Currently there are many different approaches of solving the classification problem. Many of them (for example, method SVM [1], AdaBoost [2] are focused on cases, when the number of classes l equals 2. Many solve the classification task with $l > 2$ directly (decision trees [3], artificial neural network algorithms [4]. The main idea of ECOC [5] is following. A binary (coding) matrix is formed. The matrix contains l strings (coding strings), each of which corresponds to some class. All strings of the matrix are different. N rows of the matrix define two "macro-classes," each of which is the union of some of the initial classes. For each column a binary classification algorithm is constructed.

During the classification of a new object each of N algorithms is independently applied and a binary codeword length of N is calculated. Finally, object is referred to that class, which codeword has a minimum Hamming distance to object's codeword.

[1] This paper uses the materials of the report submitted at the 9th Open German-Russian Workshop on Pattern Recognition and Image Understanding, held in Koblenz, December 1–5, 2014 (OGRW-9-2014).

[2] The article is published in the original.

Received June 21, 2015

A mixed coding strategy is proposed. The initial coding matrix is formed as a mixture of random columns and the columns obtained as solutions of special optimization tasks. An instance-based supervised classification approach [6] is used. The optimal columns and classification algorithm are found effectively in leave-one-out scheme.

The structure of the paper is following. Original notations, formulation of the problem and used basic instance-based supervised classification model are presented in Section 2. The optimization problems are presented in Section 3. Section 4 illustrates the approach on practical tasks. Conclusions are presented in Section 5.

2. INITIAL NOTATIONS AND PROBLEM STATEMENT

Consider the following standard problem recognition by precedents. Let there is a set M of objects $\mathbf{x}$, defined with their feature descriptions. For simplicity assume that $\mathbf{x} \in R^n$. The set is $M = \bigcup_{i=1}^{l} K_i$, $K_i \cap K_j = \varnothing$, $i \neq j$, with classes K_i, $i = 1, 2, \dots, l$. Information of this partition is given by training sample $X = \{\mathbf{x}_i, i = 1, 2, \dots, m\}$, which consists of representatives of each class: $X = \bigcup_{i=1}^{l} K_i^*$, $K_i^* \subset K_i$, $|K_i^*| = n_i$. The task is assigning $\forall \mathbf{x} \in R^n$ to one of the classes.

Let's describe Error-Correcting Output-Coding approach. Boolean matrix $\boldsymbol{\alpha} = \|\alpha_{ij}\|_{l \times N}$ is formed,

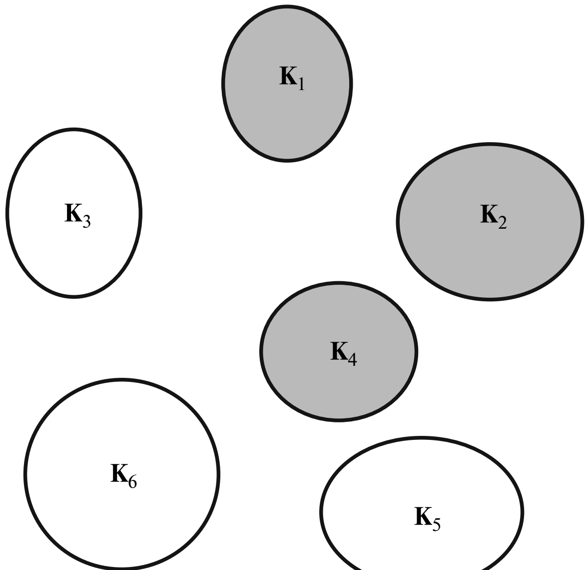

Fig. 1. Separable dichotomous task.

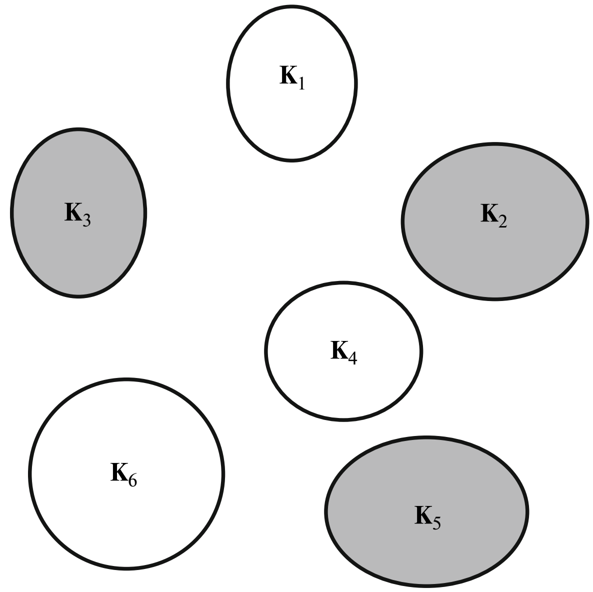

Fig. 2. Inseparable dichotomous task.

where an arbitrary column $(\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{lj})^T$, $\alpha_{ij} \in \{0,1\}, 1 < |\boldsymbol{\alpha}_j| < 0$, sets the recognition task with two classes K_1^j, K_2^j, defined by sets $\mathrm{K}_1^{j*} \subseteq \mathrm{K}_1^j$, $\mathrm{K}_2^{j*} \subseteq \mathrm{K}_2^j$, where $\mathrm{K}_1^{j*} = \bigcup\limits_{\substack{i=1,2,\ldots,l \\ \alpha_{ij}=1}} K_i^*$, $\mathrm{K}_2^{j*} = \bigcup\limits_{\substack{i=1,2,\ldots,l \\ \alpha_{ij}=0}} K_i^*$. Different columns determine various dichotomous tasks. It is assumed that all the rows of the matrix $\boldsymbol{\alpha}$ are different.

Suppose a classification model is chosen. For each pair of sets $\{\mathrm{K}_1^{j*}, \mathrm{K}_2^{j*}\}$ recognition algorithm A^j is built. Denote the result of its work with the recognition of arbitrary $\mathbf{x}$ as $A^j(\mathbf{x}) = \beta_j$. Here $\beta_j \in \{0,1\}$ means classification $\mathbf{x}$ to class K_2^j or K_1^j, respectively. Let the recognition task of $\mathbf{x}$ is solved by each algorithm A^j and boolean vector of results $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N) \in \{0,1\}$ is obtained. Object $\mathbf{x}$ refers to class K_t, $t = 1, 2, \ldots, l$, if $t = \arg\min\limits_{i=1,2,\ldots,l} \sum\limits_{j=1}^{N} |\alpha_{ij} - \beta_j|$.

The main task, which is set in the paper is optimization of α matrix, initially chosen randomly. Clearly, the columns should be not only separable, but also provide high recognition accuracy in the corresponding problem. Figures 1 and 2 show the cases of "separable" and "inseparable" pair formation $\{\mathrm{K}_1^{j*}, \mathrm{K}_2^{j*}\}$.

Construct such tasks $\{\mathrm{K}_1^{j*}, \mathrm{K}_2^{j*}\}$ (and their corresponding columns in $\boldsymbol{\alpha}$) in which their corresponding dichotomous recognition algorithm will have the best possible quality. In this case one has to calculate the quality of multiple recognition algorithm for different pairs $\{\mathrm{K}_1^{j*}, \mathrm{K}_2^{j*}\}$, therefore crucial issue is to quickly assess the quality of recognition algorithm. In this paper, as a basic model recognition model is used estimation calculation algorithm [6], for which it is possible to quickly estimate the quality of the algorithm. Fix a natural $k, 1 \le k \le n$, training set $X = \{\mathbf{x}_i, i = 1, 2, \ldots, m\}$ and the partition of it into l classes. Recognition algorithm performs a comparison of the object $\mathbf{x}$ with all objects of the training sample over all sets $\Omega \subseteq \{1, 2, \ldots, n\}$, $|\Omega| = k$ with k features and calculates a "degree of membership" (evaluation) of the object to each of the classes:

$\Gamma_i(\mathbf{x}) = \dfrac{1}{n_i} \sum\limits_{\mathbf{x}_t \in K_i} \sum\limits_{\Omega: |\Omega| = k} B_\Omega(\mathbf{x}_t, \mathbf{x})$, where the proximity function of the pair of objects is defined as $B_\Omega(\mathbf{x}_t, \mathbf{x}) = \begin{cases} 1, & |x_{tj} - x_j| \le \varepsilon_j, \quad \forall j \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$

In [6] is shown, that $\Gamma_i(\mathbf{x}) = \dfrac{1}{n_i} \sum\limits_{\mathbf{x}_t \in K_i} C^k_{d(\mathbf{x}_t, \mathbf{x})}$, where $d(\mathbf{x}_t, \mathbf{x}) = |\{j : |x_{tj} - x_j| \le \varepsilon_j, j = 1, 2, \ldots, n\}|$. Here $\varepsilon_j, j = 1, 2, \ldots, n$-parameters. Usually, they are set as $\varepsilon_j = \dfrac{2}{m(m-1)} \sum\limits_{\substack{u,v=1,2,\ldots,m \\ u>v}} |x_{uj} - x_{vj}|, j = 1, 2, \ldots, n$. After calculating the values $\Gamma_i(\mathbf{x}), i = 1, 2, \ldots, l$ object $\mathbf{x}$ refers to the class K_j with a maximum vote: $\Gamma_j(\mathbf{x}) > \Gamma_i(\mathbf{x})$, $i, j = 1, 2, \ldots, l$, $i \ne j$. Otherwise, a failure of recognition $\mathbf{x}$ occurs.

Table 1. Basic parameters of databases

Database	l	n	m	$n_i^{\min}$	$n_i^{\max}$	Attributes
Card	10	21	2126	53	579	Real
Letter	26	16	1981	62	93	Real
LS	15	35	290	10	40	Nominal, ordered
Yeast	9	8	1484	5	463	Real

3. CONSTRUCTION OF OPTIMAL DICHOTOMIES

Let the following values are precalculated: $d_{ij} = C^k_{d(\mathbf{x}_i,\mathbf{x}_j)}$, $i,\ j = 1, 2, \ldots, m,\ i \neq j$,

$$G_t(\mathbf{x}_i) = \sum_{\substack{\mathbf{x}_j \in K_t \\ \mathbf{x}_j \neq \mathbf{x}_i}} d_{ij}, \quad n_t(\mathbf{x}_i) = \begin{cases} n_t - 1, & \mathbf{x}_i \in K_t, \\ n_t, & \mathbf{x}_i \notin K_t. \end{cases}$$

Let $\mathbf{y} = (y_1, y_2, \ldots, y_l)$, $y_i \in \{0,1\}$, $l > |\mathbf{y}| > 0$—vector of variables. Define a partition of the training sample X into two classes as following:

$$\mathrm{K}_1^* = \bigcup_{\substack{i=1,2,..,l \\ y_i=1}} K_i^*, \mathrm{K}_2^* = \bigcup_{\substack{i=1,2,..,l \\ y_i=0}} K_i^*.$$

As a criterion for assessing the accuracy of the recognition with estimation calculation algorithm, corresponding the training set X and classes $\{\mathrm{K}_1, \mathrm{K}_2\}$, take the assessment of the probability of correct recognition share a recognition of objects in the leave-one-out mode.

For each training object following assessments for classes are calculated:

$$\Gamma_1(\mathbf{x}_i) = \frac{1}{\sum_{t:y_t=1} n_t(\mathbf{x}_i)} \sum_{t:y_t=1} G_t(\mathbf{x}_i)$$

and

$$\Gamma_2(\mathbf{x}_i) = \frac{1}{\sum_{t:y_t=0} n_t(\mathbf{x}_i)} \sum_{t:y_t=0} G_t(\mathbf{x}_i).$$

Quality criterion is calculated as:

$$\Phi(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} \theta(\mathbf{x}_i), \qquad (1)$$

where

$$\theta(\mathbf{x}_i) = \begin{cases} 1, & (\mathbf{x}_i \in K_t) \wedge ((\Gamma_1(\mathbf{x}_i) > \Gamma_2(\mathbf{x}_i)) \wedge (y_t = 1) \\ & \vee (\Gamma_2(\mathbf{x}_i) > \Gamma_1(\mathbf{x}_i)) \wedge (y_t = 0)), \\ 0, & \text{otherwise.} \end{cases}$$

Note that criterion (1) is computed efficiently (polynomial) for any admissible $\mathbf{y}$. The task of finding an optimal dichotomy consists of solving the following discrete optimization problem (2)–(3)

$$\Phi(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} \theta(\mathbf{x}_i) \xrightarrow[\mathbf{y}]{} \max, \qquad (2)$$

$$\mathbf{y} = (y_1, y_2, \ldots, y_l), \quad y_i \in \{0,1\}, \quad l > \sum_{i=1}^{l} y_i > 0. \qquad (3)$$

Choosing different initial approximations $\mathbf{y}$ (initial dichotomies $\{\mathrm{K}_1, \mathrm{K}_2\}$) and solving (2)–(3) problem we obtain the final matrix $\boldsymbol{\alpha} = \|\alpha_{ij}\|_{l \times N}$ of dichotomies.

For each column i, we have estimation P_i of probability of correct classification in corresponding dichotomy. Values P_i are calculated according to (1). Then the approximate minimization of the matrix $\boldsymbol{\alpha}$ is to remove all columns for which $P_i < \lambda \frac{\sum_{j=1}^{N} P_j}{N}$ provided that all the rows of the reduced matrix will be different, $0 < \lambda \leq 1$.

4. RESULTS OF NUMERICAL EXPERIMENTS

The repository [7] has been used for testing and comparing algorithms over multi-class problems. These were the following tasks.

Task “Cardiotocography” or “card” (briefly). The dataset consists of measurements of fetal heart rate and uterine contraction features on cardiotocograms classified by expert obstetricians. Task “Letter Image Recognition Data” or “letter.” The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. Task “Large Soybean” (“LS”) is classification of soybean disease. The last task was recognition of protein localization sites (task “yeast”).

Parameters $n_i^{\min}$ and $n_i^{\max}$ indicate the minimum and maximum numbers of the objects in the classes of corresponding database. All major parameters of databases are listed in Table 1.

The results of the experiments are shown in Table 2.

Initial number of binary columns was 1000. The column DA (direct application) shows the results of the direct application of an algorithm for the case of l classes. ECOC column shows the results of classifica-

Table 2. The results of comparative experiments

Database	k	λ	N_{final}	ECOC, %	DA, %
Card	8	1	277	70.69	61.10
Letter	6	1	157	62.14	58.66
LS	10	1	151	87.93	87.24
Yeast	5	0.8	153	58.83	56.33

tion using the approach "Error-Correcting Output Codes." As an evaluation of the quality, was used the percentage of correct answers in leave-one-out mode.

5. CONCLUSION

In this paper a problem of optimizing the code table when applying ECOC in the supervised classification problem with many classes was considered. The main problem in this model is the problem of construction matrix $\boldsymbol{\alpha}$. This problem is reduced to a certain integer programming problem. It was implemented a greedy algorithm to solve it. Results of preliminary experiments on different databases are presented in Table 2.

REFERENCES

1. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
2. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci. **55** (1), 119–139 (1997).
3. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, 1993).
4. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychol. Rev. **65** (6), 386–408 (1958).
5. T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," J. Artificial Intelligence Res. **2**, 263–282 (1995).
6. Yu. Zhuravlev, *Selected Publications* (Magistr, Moscow, 1998) [in Russian].
7. K. Bache and M. Lichman, *UCI Machine Learning Repository* (School of Information and Computer Science, Univ. of California, Irvine, CA, 2013). http://archive.ics.uci.edu/ml

Vasilii Vladimirovich Ryazanov, has master degree in applied mathematics and physics. Graduated from the Moscow Institute of Physics and Technology (specialty "Computer science") in 2014. Post-graduate student and assistant at Computer Science department at Moscow Institute of Physics and Technology. He is an author of 5 scientific articles. Fields of scientific interests: machine learning, data mining, econometrics, mathematical problems of recognition, classification and forecasting, python programming, web development.