System Evaluation of Error Correcting Output Codes for Artificial Data Models

Shigeichi Hirasaw [a]****, Gendo Kumoi*** [b]***, Manabu Kobayash*** [c]***, Masayuki Goto*** [d]***, and Hiroshige Inazumi*** [e]

[a] *Research Institute for Science and Engineering, Waseda University*
3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555 JAPAN
hira@waseda.jp
[b] *Graduate School of Creative Science and Engineering, Waseda University*
3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555 JAPAN
moto-aries@ruri.waseda.jp
[c] *Center for Data Science, Waseda University*
1-1-4, Totsuka, Shinjuku, Tokyo 169-8050 JAPAN
mkoba@waseda.jp
[d] *School of Creative Science and Engineering, Waseda University*
3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555 JAPAN JAPAN
masagoto@waseda.jp
[e] *Faculty of Informatics, Aoyama Gakuin University*
5-10-1, Fuchinobe, Chuo, Sagamihara, Kanagawa 252-5258 JAPAN
inazumi@si.aoyama.ac.jp

**Corresponding Author: hira@waseda.jp*

ABSTRACT

Performance of Multi-level classification systems constituted by a plural number of binary classifiers usually called ECOC (Error Correcting Output Codes) is discussed and evaluated, assuming the M-dimensional Normal Distribution for a classification data model with M (≥ 3) categories. First, based on this artificial model, the relationship between the number of binary classifiers N and the classification error probability P_e is investigated, and easily found it to be in trade-offs. Starting with the exhaustive codes of code length $N_{\max}=2^{M-1}-1$, we clarify the performance of shortened version of the exhaustive codes of length N ($\leq N_{\max}$). Here, the average performance for P_e is discussed, and note that it is not the object of this study to obtain the construction method of the ECOC which minimizes P_e. Next, we regard the two variables, N and P_e which are in a trade-off relationship, N as the investment cost and P_e as the performance degradation, and normalize both of them with their maximum values. That is, we let $n=N/N_{\max}$, and $p_e=P_e/P_{e,\max}$, where $P_{e,\max}$ is the value of P_e when $N=M-1$. Letting the number of categories M as a parameter which gives the scale of the system, we apply them to the system evaluation model in the OR fields. As the result, the system trade-off functions between n and p_e are shown to have desirable properties, such as "Flexible" and “Elastic”. Here, “Flexible” means that the system has a downward convex and decreasing function, hence this suggests that we can decrease the investment cost drastically with tolerating a slight increase in the performance degradation. While “Elastic” implies that the system has a function which approaches to origin as M becomes large. Hence the ECOC has desirable condition as the number of categories M becomes large. Note that these results are similarly obtained when applied to real data such as document classification and hand-written character recognition tasks.

Keywords: Multi-level classification, Binary classifier, Trade-off model, System evaluation model, ECOC, Exhaustive code, Error correcting codes, Artificial data model

1. INTRODUCTION

In the field of machine learning, there are many multi-level classification (multi-class) problems such as the document classification [19], and the hand-written character recognition [17]. Although there is a method for directly solving the multi-class problems using a single multi-level classifier, it is generally not practical because of computational complexity. In the present study, we consider the multi-class problems using multiple binary classifiers which has been known and studied as the ECOC (Error Correcting Output Codes) [1][2][5][14]. We use SVMs (Support Vector Machines) and the RVMs (Relevance Vector Machines) which are known to perform well as binary classifiers.

On the other hand, J. Pearl and A. Crolotte discussed the trade-off between the amount of memory and the error in QA (Question Answering) systems based on rate-distortion theory [15]. They clarified the conditions such that we can reduce the large amount of memory, if the small error rate can be tolerated. They introduced desirable conditions for systems such as “Flexible" and “Elastic". In our previous work, we have applied this theoretical model to various tasks [11]. However, it imposes some strong restrictions to target information systems, since the model is based on rate-distortion theory. Subsequently we successfully removed these restrictions and also extended the desirable conditions to make them useful by generalized trade-off model used for system evaluation [7][8], which is a kind of the trade-off model as seen in the OR (Operations Research) area. It would be useful for evaluation of information systems prior to start researching, developing, or designing them.

In this paper, we apply the trade-off model for system evaluation to construction methods of the multi-class systems using binary classifiers, assuming the data be generated by artificial model [3][4][16][18]. We discuss on the trade-off between the number of binary classifiers (*investment cost*) N and the probability of classification error (*performance degradation*) P_e of the multi-class system configuration with the number of categories (*scale of the system*) M as a parameter. Then we investigate whether the systems satisfy desirable conditions or not as M increases. It should be noted that minimizing the probability of classification error [13] is not the objective of this study.

Throughout this paper, we shall evaluate the *average* performance of the construction methods which solve the multi-class problems using binary classifiers. In section 2, we briefly describe the configuration methods. Section 3 shows the construction methods of the code word table. Experiments and discussions are described in sections 4 and 5, respectively. Finally section 6 gives concluding remarks. The trade-off model for system evaluation called the system evaluation model is summarized in Appendix A.

2. MULTI-LEVEL CLASSIFICATION SYSTEM USING BINARY CLASSIFIERS

The main part of the configuration for the multi-level classification system using binary classifiers is usually called the ECOC Matrix [12], the classifier structure, the code word structure, the coding matrix and so on. Here we call it a *code word table.*

2.1. Multi-class System Configuration

The multi-class system configuration using binary classifiers is shown in Figure 1. The main part of this configuration is the code word table. The rest of it is the binary classifier.

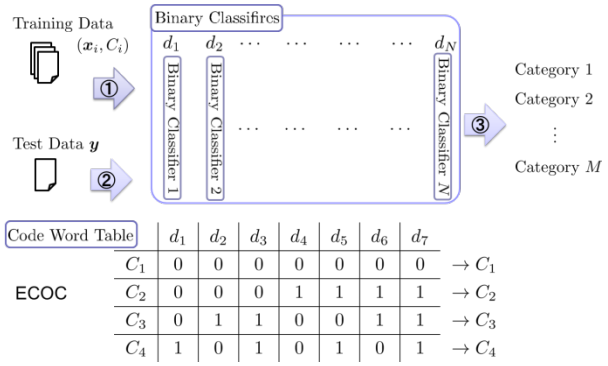

Figure 1. Multi-class system configuration using binary classiers.

2.2. Code Word Table

The code word table shown in Figure 2 is represented by the matrix as follows:

$$W = [w_{ij}] \quad (w_{ij} \in \{0, 1\},\ i = 1,\ 2,\ \cdots, M,\ j = 1,\ 2, \cdots,\ N)$$
$$= [\boldsymbol{d}_1{}^{\mathrm{T}} 1, \boldsymbol{d}_2{}^{\mathrm{T}}, \cdots, \boldsymbol{d}_N{}^{\mathrm{T}}]$$
$$= [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M]^{\mathrm{T}} \qquad (1)$$

where T represents the transpose of a matrix (or a vector).

[Example 1] Table 1 shows the case where $M = 5$, $N = 5$, which is called "one vs. the rest" method and it is one of basic types of the code word table.

Table 1. Example of code word table of "one vs. the rest" method (M=5, N=5, D=2).

	d_1	d_2	d_3	d_4	d_5
C_1	1	0	0	0	0
C_2	0	1	0	0	0
C_3	0	0	1	0	0
C_4	0	0	0	1	0
C_5	0	0	0	0	1

2.3. Binary Classifier

For learning from the examples, the training data $\boldsymbol{x}$'s are given in the form of $(\boldsymbol{x};\ C_i)$, where $C_i \in \boldsymbol{C}$ represents the i-th category and $\boldsymbol{C}$ is a set of categories. Using the function $f(\cdot)$ learned from the training data, the test data (whose category is unknown) $\boldsymbol{y}$ is classified into $C_{i'}$ estimated by $f(\boldsymbol{y}) = C_{i'} \in \boldsymbol{C}$, where

$$f(\boldsymbol{y}) = \arg\max_{C_i \in \boldsymbol{C}} g_i(\boldsymbol{y}). \qquad (2)$$

The j-th binary classifier d_j of the maximum margin soft decision SVM[1][6] outputs $h_j(\boldsymbol{y}) \in (-\infty, +\infty)$ to calculate $g_i(\boldsymbol{y}) = \Sigma_j\, I(w_{ij})\, h_j(\boldsymbol{y})$, where $I(w_{ij})$ takes 1 for $w_{ij} = 1$, and -1 for $w_{ij} = 0$.

2.4. System Evaluation Model

The object system modeled in subsection 2.1 is applied to the trade-off model described in Appendix A. Experimental data are fed into the part of the multi-class system configuration, and the probability of classification error P_e is obtained as the performance degradation of the system for the number of binary classifiers N as a variable. Here, the complexity of the problem, that is, the scale of the system is given by the number of categories M to be classified. Table 2 shows the correspondence between variables and parameters of rate-distortion theory, those of the trade-off model for system evaluation, and those of the object system i.e., the multi-class system.

Table 2. Evaluation of Multi-class System (Correspondence Table) [1]

Rate-Distortion Theory	Trade-off Model	Multi-class System
Rate (L)	Investment Cost (ℓ)	Number of Binary Classifiers (n)
Distortion (D)	Performance Degradation (d)	Probability of Classification Error (p_e)
	Scale of System (G)	Number of Categories (M)

3. CONSTRUCTION OF CODE WORD TABLE

In this section, construction methods for code word table are discussed.

3.1 Generation of Exhaustive Codes

One of the most important code word table is given by the exhaustive code [5]. The code word table of the exhaustive code is generated by

(i) choose all the column vectors of length M,

(ii) remove the complement column vectors from them, and after that,

(iii) remove the all 1 (or all 0) column vector.

Consequently, the length of the (full) exhaustive code is given by:

$$N_{\max} = 2^{M-1} - 1. \tag{3}$$

[Example 2] An example of the exhaustive code with $M = 5$, and $N_{\max} = 15$ is shown in Table 3.

Table 3. Exhaustive code ($M = 5$, and $N_{\max} = 15$)

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}
C_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C_2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
C_3	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1
C_4	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
C_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

3.2 Construction of Shortened Exhaustive Codes

As is implied by the name, the exhaustive code extracts column vectors exhaustively.

[1] In the following sections, unlike Figure A.1, the horizontal axis is used for n, and the vertical axis, for p_e.

Since the code length of the (full) exhaustive code is given by the Eq. (3), let us consider a

shortened version of exhaustive code of length N with $N_{min} \leq N \leq N_{max}$, where, $N_{min} = M - 1$, i.e. , length of the "modified one vs. the rest" method [2]. Then we can decrease the investment cost (decreasing the number of binary classifiers) N by tolerating the performance degradation (increasing the probability of classification error) P_e. For given M categories, we shall evaluate the relationships between the number of binary classifiers N and the *average* probability of classification error P_e, where P_e corresponding to N column vectors selected from the N_{max} column vectors is obtained and is averaged over N column combinations out of all N_{max} columns [3]. The obtained results are normalized by N_{max} and $P_{e,max}$, and we have the normalized function with $n = N / N_{max}$, $p_e = P_e / P_{e,max}$:

$$p_e = s(n, M) \tag{4}$$

where $P_{e,max}$ corresponds to the value of P_e for $N = N_{min} = M - 1$.

4 EXPERIMENTS

Let us obtain and show the Eq. (4) defined in the previous section by experiments.

4.1 Artificial Data Generation

Consider the M-dimensional Normal Distribution $N(\mu, \Sigma)$ whose probability density function $g(\boldsymbol{z})$ is given by

$$g(\boldsymbol{z}) = \frac{1}{(2\pi)^{M/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right\} \tag{5}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_M)^{\mathrm{T}}$ and $\Sigma = [\sigma_{ij}]$ $(i, j = 1,2, \cdots, M)$. Assuming learning data $\boldsymbol{x}$, and test data $\boldsymbol{y}$ be generated by M-dimensional Normal Distribution, that is $\boldsymbol{x}, \boldsymbol{y} \sim N(\boldsymbol{\mu}, \Sigma)$.

4.2 Experiments by Artificial Data

By using random number generator with M-dimensional normal distribution, specification of learning data $\boldsymbol{x}$ and $\boldsymbol{y}$ are given as shown in Table 3.

Table 4. The number of experimental data

Number of categories M	4, 5, 6, 7, 8
Number of learning data/category	100
Number of test data/category	100
Number of trials	20

[Experiment 1] Isotropic synthetic data generation with no correlation for M=8
As the most ideal and simple case, we choose $\mu_i = 1$, $i = 1, 2, \cdots, M$, $\sigma_{ii} = 0.5$, and

[2] For the purpose of comparison, we use "modified one vs. the rest" method by removing the column vector $(0, 0, \cdots, 0, 1)^{\mathrm{T}}$ of "one vs. the rest" method as shown in Table 1.

[3] It is equivalent to randomly choose N columns among the all N_{max} columns with the uniform probability distribution.

$\sigma_{ij}=0$ (for all $i \neq j$), $i, j = 1, 2, \cdots, M$, i.e., with no correlation. The result obtained is depicted in Figure 2 together with the results obtained by real data (document classification data [19])[4]. This figure shows the relationship between the *average* probability of classification error P_e and the number of the binary classifiers N for M=8, and N_{max}=127, where the N columns are randomly chosen from N_{max} columns, and each dot indicates a result obtained by an N combination of N_{max} column vectors. A set of dots is illustrated as if it were a heavy vertical line.

[Experiment 2] Isotropic synthetic data generation with correlation for M=8

For the same conditions as experiment 1 except for $\sigma_{ij}=0.125$ (for all $i \neq j$), i.e., the coefficient of correlation is 0.5, the result obtained is also shown in Figure 2.

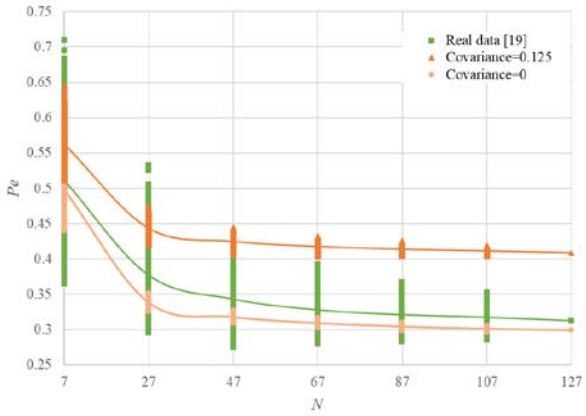

Figure 2. The relationship between the probability of classification error P_e and the number of binary classifiers N (M=8)

[Experiment 3] Comparison with the minimum distance classification method

In experiment 1 and 2, we have used the SVM as a binary classifier. If we use the minimum distance classification (MDC) method instead of the SVM, the probability of classification error P_e would become large. The result is illustrated in Figure 3 for $\mu_i = 1$, $i = 1, 2, \cdots, M$, $\sigma_{ii}=0.5$, and $\sigma_{ij}=0$ (for all $i \neq j$), $i, j = 1, 2, \cdots, M$, i.e., with no correlation.

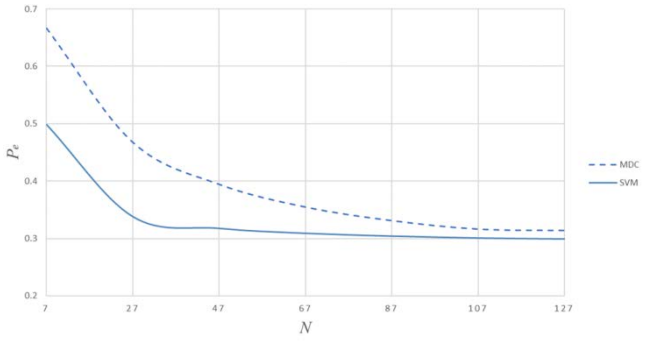

Figure 3. The relationship between the probability of classification error P_e and the number of binary classifiers N to compare the case of using the minimum distance classification (MDC) method (M=8)

[4] The result for real data set [19] is copied from Figure 5.1 in [9]

5 DISCUSSIONS

5.1. The relationship between P_e and N

The relationship between the probability of classification error P_e and the number of binary classifiers N is shown to be in trade-offs as Figure 2 and Figure 3.

[Remark 1] From Figure 2, we have:

- The case σ_{ij}=0.125(for all $i \neq j$) increases the probability of classification error P_e compared to the case σ_{ij}=0 (for all $i \neq j$) with the same σ_{ii}=1.0 for the Isotropic synthetic data**.**
- For the case of actual data set, the result for a set of N combination among N_{max} columns varies widely [5] as seen in the heavy vertical lines in Figure 3,

[Remark 2] From Figure 3, we have:

- Compared to the minimum distance classification (MDC) method for binary classifiers**,** the SVM classifier for them performs better, since the probability of classification error P_e by the latter is lower than that by the former.

5.2. The relationship between p_e and n

The trade-off curves which corresponds to the Eq. (4) obtained by these experiments are shown in Figure 4 and Figure 5. Figure 4 shows the trade-off curves for the case of M = 4, 5, 6, 7, and 8, when μ_i= 1, i = 1, 2, ・・・, M, σ_{ii}=0.5, and σ_{ij}=0 (for all $i \neq j$), i, j = 1, 2, ・・・, M, i.e., with no correlation. While figure 5 shows those for the same conditions except for using the minimum distance classification (MDC) method instead of the SVM.

Applying the target system i.e., multi-classification system constructed by binary classifiers (ECOC) to the system evaluation model described in Appendix A, we shall find the following interesting properties.

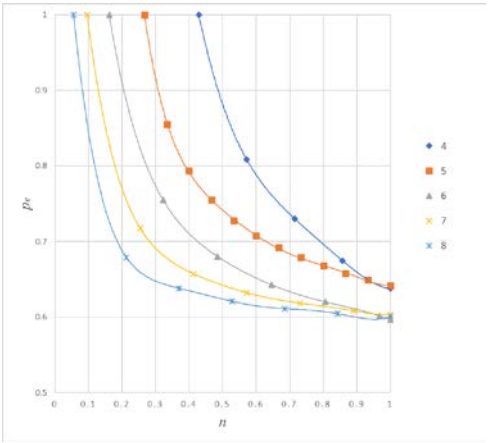

Figure 4. Trade-off relationship between investment cost n and performance degradation p_e with system scale parameter M for μ_i=1, i=1, 2, ・・・, M, and σ_{ii}=0.5, σ_{ij}=0 ($i \neq j$).

[Remark 3] From Figure 4, we have:

- The trade-off curves are decreasing and convex downward, hence are shown to be **flexible**.
- Most of the trade-off curves go toward the origin as M becomes large except in the neighborhood of n = 1, and are almost **elastic.**

[5] Actual data may be generally not isotropic

[Remark 4] From Figure 5, we have:

- For the case of using the minimum distance classification (MDC) method instead of the SVM, the similar properties to Remark 3 holds, hence the system has also **flexible** and **elastic**.

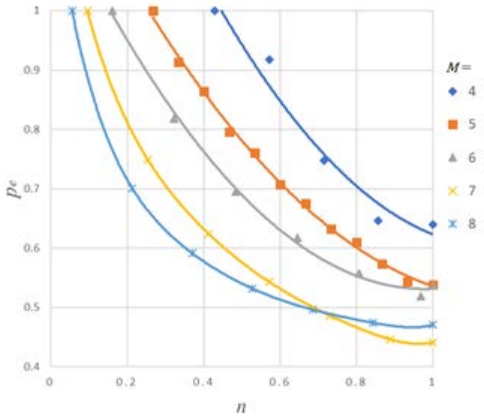

Figure 5. Trade-off relationship between investment cost n and performance degradation p_e with system scale parameter M to compare the case of using the minimum distance classification (MDC) method for μ_i=1, i=1, 2, ・・・, M, and σ_{ii}=0.5, σ_{ij}=0 ($i \neq j$).

6 CONCLUDING REMARKS

In this paper, we investigated the method for constructing multi-class systems using binary classifiers (the main part is called the ECOC), assuming the data be generated by artificial model. First, we clarify the relationship between the probability of classification error P_e and the number of binary classifiers N is in trade-offs by experiments. Next, the trade-off model used for system evaluation is applied to this result, and we evaluated the ECOC systems in terms of the trade-off relationship between the investment cost n and the performance degradation p_e. Our trade-off results show that they have desirable properties such as *flexible and elastic* when increasing the scale of the system M. Our main findings are highlighted as Remarks 1 to 4.

Although we have conducted empirical evaluations using artificial data for only isotropic conditions this time, it is also necessary to investigate in details for different values of μ and Σ, and to clarify the analytical performance evaluation of the ECOC. In the future, we would like to discuss the case where the fruits of coding theory are effectively introduced, especially we expect that the modified Reed-Muller (mRM) codes [6] would play an important role in this area.

APPENDIX

Appendix A: System Evaluation Model

Introducing rate-distortion theory, we briefly describe the trade-off model for system evaluation called (a generalized version of) system evaluation model [7][8].

A.1. Outline of Rate-Distortion Theory

Rate-distortion theory discusses data compression by the trade-off property between rate and distortion [15]. The rate-distortion function can be written as:

$$L = R(D) \tag{A.1}$$

where L is the rate defined by $L = (1/n) \log |C|$, where $|C|$ is the number of code words, n, the code length, and D, the distortion. The $L = R(D)$ is usually a convex downward and non-increasing function of D.

A.2. Trade-off model for system evaluation

Generally, the rate L discussed in the previous subsection corresponds to the investment cost of a system, and distortion D, the performance degradation of the system [15]. By extending the rate-distortion model, we have proposed the trade-off model for system evaluation [7][8], where we have also introduced a parameter G as the scale of the system.

Let the rate L be normalized by the maximum of L, L_{max}, and the distortion D, by the maximum of D, D_{max}, then we have the following normalized function by $\ell = L/L_{max}$, and $d = D/D_{max}$, and introducing G:

$$\ell = r(d;G) \tag{A.2}$$

For evaluation of the systems, we define the following properties to the *normalized* trade-off system evaluation function (A.2):

[Definition A.1]

(1) Flexible [15]: The system is "flexible", if $\ell = r(d;G)$ is a decreasing and convex downward function. And the system A with $\ell = r_A(d;G)$ is *more flexible* than the system B with $\ell = r_B(d;G)$, if $r_A(d;G) < r_B(d;G)$ for arbitrary d ($0 < d < 1$), and G ($G > 1$). (See Figure A.1).

(2) Elastic [15]: The system with $\ell = r(d;G)$ is *elastic*, if $\ell = r(d;G)$ is a decreasing function of G for arbitrary d ($0 < d < 1$). (See Figure A.1).

(3) Effective elastic [7]: The system is *effective elastic*, if the system is elastic and $\ell = r(d;G)$ is a convex downward function of G.

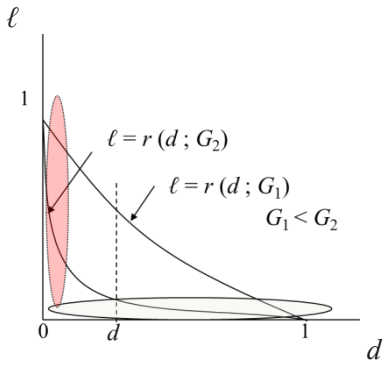

Figure A.1. Trade-off model.

As shown in Figure A.1, ℓ is a decreasing and convex downward function of d, hence we can decrease ℓ drastically tolerating a slight increase in d.

REFERENCES

[1] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," Journal of Machine Learning Research, vol.1, pp.113-141, 2000.

[2] G. Armano, C. Chira, and N. Hatami, "Error-correcting output codes for multi-label text categorization," Proceedings of the Third Italian Information Retrieval Workshop, IIR 2012, pp.26-37, Italy, Jan. 2012.

[3] L. Cai, T. Hofmann, "Hierarchical document categorization with support vector machines," the 13th ACM International Conference on Information and Knowledge Management (CIKM' 04), November 8–13, 2004, Washington, DC, USA, Nov., 2004.

[4] E. J. Coyle, R. G. Roberts, E. G. CollinsJr., and A. Barbu, "Synthetic data generation for

classification via uni-modal cluster interpolation," Autonomous Robots, vol. 37, no.1 pp.27–45, June 2014.
[5] T. G. Dietterich and G. Bakiri: "Solving multi-class learning problems via error-correcting output codes," Journal of Artificial Intelligence Research, vol.2, pp.263-286, 1995.
[6] M. Goto, and M. Kobayashi, Introduction to Pattern Recognition and Machine Learning (in Japanese), Corona-Sha, Tokyo, 2014.
[7] S. Hirasawa, and H. Inazumi,"A system evaluation model by using information theory," The 30th Joint National Meeting, ORSA/TIMS, MB35.3, Philadelphia, PE. USA, Oct. 1990.
[8] S. Hirasawa, and H. Inazumi, "A model for system evaluation based on information theory," The 2000 International Conference of Management Science and Decision Making, Tamkang University, Taipei, ROC, June 2000.
[9] S. Hirasawa, G. Kumoi, M. Kobayashi, M. Goto, and H. Inazumi, "A System evaluation of construction methods for multi-class problems using binary classifiers," (in Japanese), Proceedings of the 2016 Fall Conference, The Japan Society for Management Information (JASMIN), B1-2, Osaka, Japan, September 14-15, 2016.
[10] S. Hirasawa, G. Kumoi, M. Kobayashi, M. Goto, and H. Inazumi, "System evaluation of construction methods for multi-class problems using binary classifiers," Proceedings of the World CIST''2018, Vol. 2, pp.909-919, Napoli, Italia, March 27-29, 2018.
[11] H. Inazumi, Studies on the Evaluations for Information Systems based on Rate Distortion Theory, Dissertation of Dr. Eng, Waseda University, November 1989.
[12] N. Japkowicz, V. Barnabe-Lortie, S. Horvatic, and J. Zhou, "Multi-class learning using data driven ECOC with deep search and re-balancing," 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp.19-21, Paris, France, Oct. 2015.
[13] G. Kumoi, M. Kobayashi, M. Goto, and S. Hirasawa, "A consideration in code word composition in multi-level document classification by ECOC method (in Japanese)," FIT 2016, 15th Information Science and Technology Forum, Toyama, September 7-9, 2016.
[14] Y. Luo, and K. Najarian, \Employing decoding of specific error correcting codes as a new classification criterion in multiclass learning problems," Proceedings of 2010 International Conference on Pattern Recognition, pp.4238-4241, 2010.
[15] J. Pearl, and A. Crolotte, "Storage space versus validity of answers in probabilistic question answering systems," IEEE Trans. Inform. Theory, vol. IT-26, no. 6, pp.633-640, Nov. 1979.
[16] J. Sánchez-Monedero, P.A. Gutiérrez, M. Pérez-Ortiz, and C. Hervás-Martínez, "An n-spheres based synthetic data generator for supervised classification," International Work Conference on Artificial Neural Networks 2013 (IWANN 2013), Part 1, LNCS 7902, pp.613-621, Canary Islands (Tenerife-Puerto de la Cruz), Spain, 12-14 June 2013.
[17] UCI machine learning repository, URL: http://www.tri_elds.jp/uci-machine-learning-repository-dataset s-956G.
[18] C. van der Walt and E. Barnard, "Data characteristics that determine classifier performance," Proceedings of 16th Annual Symposium of the Pattern Recognition, Association of South Africa (SPR of SA), pp 160-165, 2006.
[19] Yomiuri News Paper Articles 2000, Naigai Associates Inc.

ACKNOWLEDGMENTS

One of the authors S. H. would like to thank Professor Shin' ichi Oishi of Waseda University for giving a chance to study this work. The authors would like to thank to Professors Hideki Yagi of Electro-Communication University, and Kenta Mikawa of Shonan Institute of Technology for their helpful suggestions for this research. The research leading to this paper was partially supported by MEXT Kakenhi under Grant-in Aids for Scientific Research (B) No. 26282090, (C) No. 16K00491 and (C) No. 18K11585.

ABOUT THE AUTHORS

Shigeichi HIRASAWA was born in Kobe, JAPAN on Oct. 2nd, 1938. He received his BS degree in Mathematics, and BE degree in Electric Communication Engineering, both from Waseda University in 1961 and 1963, respectively, and his Dr.E degree in Communication Engineering from Osaka University in 1975. He was a technical staff at Mitsubishi Electric Corporation from 1963 through 1981. He joined Waseda University as a professor in 1981, and from 2009 he has been a research consultant at Research Institute for Science and Engineering, Waseda University. His interest includes Information theory, coding theory and their applications. He is a Life Fellow of IEEE, a Fellow of IEICE, and a member of the Information Processing Society of Japan.

Gendo KUMOI received his BE degree from Waseda University in 2008. He has been an adjunct researcher, Research Institute for Science and Engineering, Waseda University form 2008, and is now a doctoral student of Graduate School of Science and Engineering, Waseda University. His research areas are data mining, machine learning, big data analysis, and coding theory. He is a member of the Information Processing Society of Japan.

Mababu KOBAYASHI was born in Yokohama, JAPAN on Oct. 30th, 1971. He received the B.E. degree, M.E. degree and Dr.E. degree in Industrial and Management Systems Engineering from Waseda University, Tokyo, Japan, in 1994, 1996 and 2000, respectively. From 1998 to 2001, he was a research associate in Industrial and Management Systems Engineering at Waseda University. From 2002 to 2018, he was a faculty of the Department of Information Science at Shonan Institute of Technology, Kanagawa, Japan. He is currently a professor of the Center for Data Science at Waseda University, Tokyo, Japan. His research interests are machine learning theory and information theory. He is a member of the Society of Information Theory and Its Applications, Information Processing Society of Japan and IEEE.

Masayuki GOTO was born in Tokyo, JAPAN, on Jan.1st, 1969. He received his B.E. and M.E. degrees from Musashi Institute of Technology, in 1992 and 1994,respectively. He received Dr.E degree in Industrial and Management Systems Engineering from Waseda University in 2000. From 2000 to 2002, he was a research associate at the graduate school of engineering, the University of Tokyo. From 2002 to 2008, he was an associate professor at Faculty of Environmental and Information Studies, Musashi Institute of Technology. From 2008, he was an associate professor at the department of Industrial and Management Systems Engineering, School of Creative Science and Engineering, Waseda University. From 2011, he is now a professor at Waseda University. He is studying in the research fields of applied information mathematics, business analytics, data science, and machine learning and its applications.

Hiroshige INAZUMI received his BE, ME, and Dr.E degrees, all in Industrial Engineering from Waseda University, in 1982, 1984, and 1989, respectively. He joined Sagami Institute of Technology as an associate professor in 1990, moved to Aoyama Gakuin University, and is now a professor. From 2016 through 2018, he was a dean of Faculty of Informatics. His research areas are information theory, artificial intelligence, machine learning, and education of Japanese language.