https://doi.org/10.20965/jaciii.2017.p0813

Paper:

Feature Selection Algorithm Considering Trial and Individual Differences for Machine Learning of Human Activity Recognition

Yuto Omae* **and Hirotaka Takahashi****

*Department of Electrical Engineering, National Institute of Technology, Tokyo College
1220-2 Kunugida, Hachioji, Tokyo 193-0942, Japan
E-mail: yuto.omae@gmail.com,
**Department of Information and Management Systems Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata 940-2188, Japan
E-mail: hirotaka@kjs.nagaokaut.ac.jp

[Received January 15, 2017; accepted July 4, 2017]

In recent years, many studies have been performed on the automatic classification of human body motions based on inertia sensor data using a combination of inertia sensors and machine learning; training data is necessary where sensor data and human body motions correspond to one another. It can be difficult to conduct experiments involving a large number of subjects over an extended time period, because of concern for the fatigue or injury of subjects. Many studies, therefore, allow a small number of subjects to perform repeated body motions subject to classification, to acquire data on which to build training data. Any classifiers constructed using such training data will have some problems associated with generalization errors caused by individual and trial differences. In order to suppress such generalization errors, feature spaces must be obtained that are less likely to generate generalization errors due to individual and trial differences. To obtain such feature spaces, we require indices to evaluate the likelihood of the feature spaces generating generalization errors due to individual and trial errors. This paper, therefore, aims to devise such evaluation indices from the perspectives. The evaluation indices we propose in this paper can be obtained by first constructing acquired data probability distributions that represent individual and trial differences, and then using such probability distributions to calculate any risks of generating generalization errors. We have verified the effectiveness of the proposed evaluation method by applying it to sensor data for butterfly and breaststroke swimming. For the purpose of comparison, we have also applied a few available existing evaluation methods. We have constructed classifiers for butterfly and breaststroke swimming by applying a support vector machine to the feature spaces obtained by the proposed and existing methods. Based on the accuracy verification we conducted with test data, we found that the proposed method produced significantly higher F-measure than the existing methods. This proves that the use of the proposed evaluation indices enables us to obtain a feature space that is less likely to generate generalization errors due to individual and trial differences.

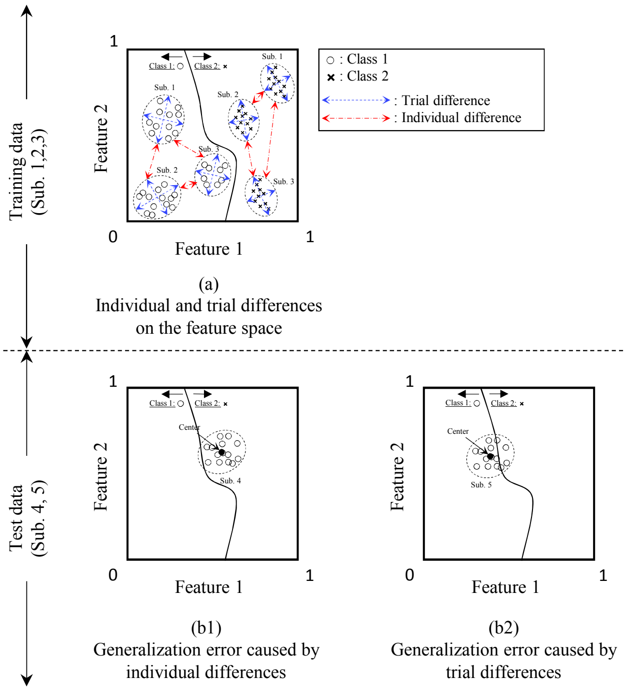
Keywords: feature selection, lower dimensional, machine learning, body motion classification, inertia sensor

1. Introduction

In recent years, many studies have been performed on the automatic determination of human body motions using data acquired from inertia sensors built into smart phones or wristband-type health care terminal devices.

Plenty of training data should be available to construct a body motion classifier. Experiments on acquiring training data should not be too long in duration out of concern for the fatigue or injury of the experimental subjects. If body motions subject to classification require a certain level of skill, it may sometimes be difficult to secure a sufficient number of subjects for the experiments. Under such circumstances, it is often difficult to acquire data from a large number of subjects. For practical reasons in many studies, therefore, a limited number of subjects repeat the same motion several times to produce the data used to construct a classifier.

For instance, Khan et al. [1] used a neural network (NN) to classify such motions as walking, running, sitting down, and standing up using sensor data acquired from six subjects; Lester et al. [2] used a hidden Markov model (HMM) to classify such motions as walking, running, brushing teeth, and riding an elevator using sensor data acquired from 12 subjects; He et al. [3] used a support vector machine (SVM) to classify such motions as walking, running, and jumping using sensor data acquired from 11 subjects; Ward et al. [4] used HMM to classify such motions as assembly work (using drills and vises) using sensor data acquired from five subjects; Siirtola et al. [5] classified motions such as swimming strokes (backstroke, crawl, turn, etc.) using sensor data acquired from 11 subjects; Kon et al. [6] classified swimming strokes



(a)
Individual and trial differences
on the feature space

(b1)
Generalization error caused by
individual differences

(b2)
Generalization error caused by
trial differences

Fig. 1. Generalization errors caused by individual and trial differences.

using a decision tree (DT) with sensor data acquired from one subject.

Classifiers constructed with data acquired from a limited number of subjects may experience problems due to individual and trial differences. Such classifiers may be unable to properly classify any unknown third party data. **Fig. 1** shows how generalization errors are caused by individual and trial differences. Classes 1 and 2 indicate the body motions to be classified, such as walking, running, and standing still. **Fig. 1(a)** shows scatter plots on a two-dimensional feature space of the sensor data for the Class 1 and 2 motions of three subjects (Sub.1, Sub.2, Sub.3) as converted into some features: Feature 1 and Feature 2. The black solid curve in **Fig. 1(a)** indicates an assumed separation boundary in machine learning (SVM, NN, etc.) by the classification models.

The red dashed arrows in **Fig. 1(a)** indicate individual differences. The feature space in **Fig. 1(a)** shows them clustered for each subject, which can be attributed to the fact that sensor data for the same motion by different subjects may vary with the individual habits of the subjects. The red dashed arrows in **Fig. 1(a)** connect the barycentric coordinates of the difference clusters for respective subjects with one another: differences in the motions of individual subjects are displayed in the feature space. Hence, we call the red dashed arrows individual differences.

The blue dotted arrows in **Fig. 1(a)** indicate trial differences. The feature space in **Fig. 1(a)** shows slight variations in the arrows for the same motion by the same subject, which can be attributed to the fact that identical sensor data cannot be acquired for the same motion by the same human being because of many different factors, such as the condition of the subject or the status of the sensor. The same motion by the same human being generates a cluster of individual differences in the same place in the feature space. The blue dotted arrows represent the expanse of a cluster of differences as acquired from the same motion by a subject (that is, differences in different trials). Hence, we refer to the blue dotted arrows as trial differences.

The individual and trial differences seem to increase generalization errors for the following reasons: **Figs. 1(b1)** and **(b2)** show plots in the feature space of generalization errors caused by individual and trial differences, respectively, when a subject who was not engaged in learning performed Class 1 motions more than once.

Figure 1(b1) shows how generalization errors can be caused by individual differences. Assuming that there are individual differences in human motions, such differences may be plotted at a slight distance from the training data (Sub.1, Sub.2, Sub.3). When such data were plotted beyond the class separation boundary, generalization errors increased.

Figure 1(b2) shows how generalization errors can be caused by trial differences. Even if the barycentric coordinates of a subject not engaged in learning stayed within the class separation boundary, they could move past the class separation boundary because of trial differences. This would also lead to an increase in generalization errors.

Any increases in generalization errors as shown in **Figs. 1(b1)** and **(b2)** can be suppressed by the following approaches: The first approach is to collect as much training data as possible. The second approach is to utilize machine learning like SVM. This technique can generate a class separation boundary that will maximize the distance between Classes.

The first approach cannot be applied unless a sufficient number of subjects participate in the research. The second approach appears to be preferable only if a feature space that is less likely to cause generalization errors (because of individual and trial differences) is already available. Evaluation indices for features and feature spaces include: the ratio of variance between the Classes to the variance within the Classes (BW-Ratio) [7]; a method using the out-of-bag samples generated during ensemble learning (OOB) [8]; and a method using feature similarities for the same class (ReliefF) [9]. However, none of these are intended to reduce generalization errors due to individual and trial differences.

In this paper, we propose evaluation indices for feature spaces that represent the likelihood of generalization errors being caused by individual and trial differences. Based on the assumption that training data can only be collected from a small number of subjects, the proposed indices are intended to search from numerous candidates for features expected to have high generalization performance. In other words, we assume that the proposed indices can be applied in the process of constructing a lower-dimensional classifier.

2. Proposed Method

2.1. Overview

In this paper, a feature space composed of features x is expressed by $[\![x]\!]$. For example, the feature space in **Fig. 1** is expressed by $[\![\text{Feature1},\text{Feature2}]\!]$.

To allow for better understanding of the proposed method, we explain it for a simple problem: "when classifying Classes into two Classes (c_1 and c_2), constitute an optimal two-dimensional feature space $[\![a^{\text{opt1}},a^{\text{opt2}}]\!]$, $(a^{\text{opt1}},a^{\text{opt2}} \in A)$ out of a set with p features as members $A = \{a_1,\ldots,a_p\}$."

In the proposed method, in order to search an optimal feature space, we calculate the indices with which to evaluate the superiority of the feature space $[\![a_i,a_j]\!]$ $(i,j = 1,\ldots,p, i \neq j)$ composed of $a_i,a_j \in A$ in the classification problem. Then, we adopt a feature space with the best indices as the optimal feature space $[\![a^{\text{opt1}},a^{\text{opt2}}]\!]$, $(a^{\text{opt1}},a^{\text{opt2}} \in A)$.

Figure 2 outlines the evaluation indices for feature spaces. The proposed method consists of the following two phases: phase 1) for constructing the feature space $[\![a_i,a_j]\!]$ $(i,j = 1,\ldots,p, i \neq j)$ that is expressed by two arbitrary features out of p feature probability distributions of Class c_n $(n = 1,2)$, considering individual differences and trial differences; phase 2) to process misclassification likelihood into one-dimensional actual values, using the probability distributions. The phase 1) is referred to in Sections 2.2 to 2.4 and the phase 2) is referred to in Section 2.5.

In order to construct probability distributions for Class c_n that take individual and trial differences into consideration, we first construct the probability distributions that generate individual differences. Then, we construct the probability distributions that generate trial differences. In particular, the method for constructing the probability distributions that generate individual differences is described in Section 2.2, and the method for constructing the probability distributions that generate trial differences is referred to in Section 2.3.

2.2. Probability Distributions of Individual Differences

As shown by the red dashed arrows in **Fig. 1(a)**, the barycentric coordinates in the feature space may differ with the subjects. We first construct the probability density function to generate in the feature space different barycentric coordinates with the subjects and generate their barycenters by following said function. We have adopted a multivariate normal distribution as the shape of the probability density function. For the actually acquired data for all M subjects, we calculate the mean barycentric coordinate of each subject and the variance-covariance matrix; these constitute the parameters of the multivariate normal distributions.

If subject m performs Class c_n motions more than once, the mean value of a_i and a_j is denoted by $\overline{a_{i,m}^{\{c_n\}}}, \overline{a_{j,m}^{\{c_n\}}}$. By executing the same operation for all of M subjects, the mean vector $\boldsymbol{w}_{a_{i,j}}^{\{c_n\}}$ and variance-covariance matrix $V_{a_{i,j}}^{\{c_n\}}$ are calculated by the following equations:

$$\boldsymbol{w}_{a_{i,j}}^{\{c_n\}} = \left(w_{a_i}^{\{c_n\}}\ w_{a_j}^{\{c_n\}}\right)^T, \tag{1}$$

$$V_{a_{i,j}}^{\{c_n\}} = \begin{pmatrix} v_{a_i}^{\{c_n\}\ 2} & v_{a_{i,j}}^{\{c_n\}} \\ v_{a_{i,j}}^{\{c_n\}} & v_{a_j}^{\{c_n\}\ 2} \end{pmatrix}, \tag{2}$$

where $w_{a_i}^{\{c_n\}}$ denotes the mean value for all of M subject of $\overline{a_{i,m}^{\{c_n\}}}$ and is calculated by the following equation:

$$w_{a_i}^{\{c_n\}} = \frac{1}{M}\sum_{m=1}^{M}\overline{a_{i,m}^{\{c_n\}}}. \tag{3}$$

$w_{a_j}^{\{c_n\}}$ can be calculated by replacing subscript i in Eq. (3) with j (hereinafter, denotations are omitted without notice if replacement of subscripts fulfills calculations and causes no special confusion). $v_{a_i}^{\{c_n\}\ 2}$ and $v_{a_{i,j}}^{\{c_n\}}$ denote the variance and covariance of $\overline{a_{i,m}^{\{c_n\}}}$, respectively, and are calculated by the following equations:

$$v_{a_i}^{\{c_n\}\ 2} = \frac{1}{M}\sum_{m=1}^{M}\left(w_{a_i}^{\{c_n\}} - \overline{a_{i,m}^{\{c_n\}}}\right)^2, \tag{4}$$

$$v_{a_{i,j}}^{\{c_n\}} = \frac{1}{M}\sum_{m=1}^{M}\left(w_{a_i}^{\{c_n\}} - \overline{a_{i,m}^{\{c_n\}}}\right)\left(w_{a_j}^{\{c_n\}} - \overline{a_{j,m}^{\{c_n\}}}\right). \tag{5}$$

By adopting the mean value vector $\boldsymbol{w}_{a_{i,j}}^{\{c_n\}}$ and variance-covariance matrix $V_{a_{i,j}}^{\{c_n\}}$ obtained from the calculations as the parameters of the multivariate normal distributions, unknown subject m's barycentric coordinates $\boldsymbol{a}_{i,j,m'} = (a_{i,m'}\ a_{j,m'})^T$ in the feature space are generated by following the probability distributions:

$$F_{\text{Individual}}^{\{c_n\}}(\boldsymbol{a}_{i,j,m'}) = \frac{1}{(\sqrt{2\pi})^2\sqrt{|V_{a_{i,j}}^{\{c_n\}}|}} \times \exp\left(-\frac{1}{2}\left(\boldsymbol{a}_{i,j,m'} - \boldsymbol{w}_{a_{i,j}}^{\{c_n\}}\right)^T V_{a_{i,j}}^{\{c_n\}\ -1} \times \left(\boldsymbol{a}_{i,j,m'} - \boldsymbol{w}_{a_{i,j}}^{\{c_n\}}\right)\right). \tag{6}$$

The barycentric coordinates of Class c_n by M' subjects are generated in the feature space by following the probability distributions. The second feature space on left side of **Fig. 2** represents the individual barycentric coordinates of $M' = 9$ people (i.e., $m' = 1,\ldots,9$) that have been generated following $F_{\text{Individual}}^{\{c_n\}}(\boldsymbol{a}_{i,j,m'})$. Thus, the use of $F_{\text{Individual}}^{\{c_n\}}(\boldsymbol{a}_{i,j,m'})$ enables us to generate individual barycentric coordinates in the feature space.

2.3. Probability Distributions of Trial Differences

The discussions in Section 2.2 have accomplished constructing M' subjects' barycentric coordinates in the feature space. We represent trial differences by adding variance to the obtained barycentric coordinates and spreading them (i.e., the variance added to the barycentric coor-

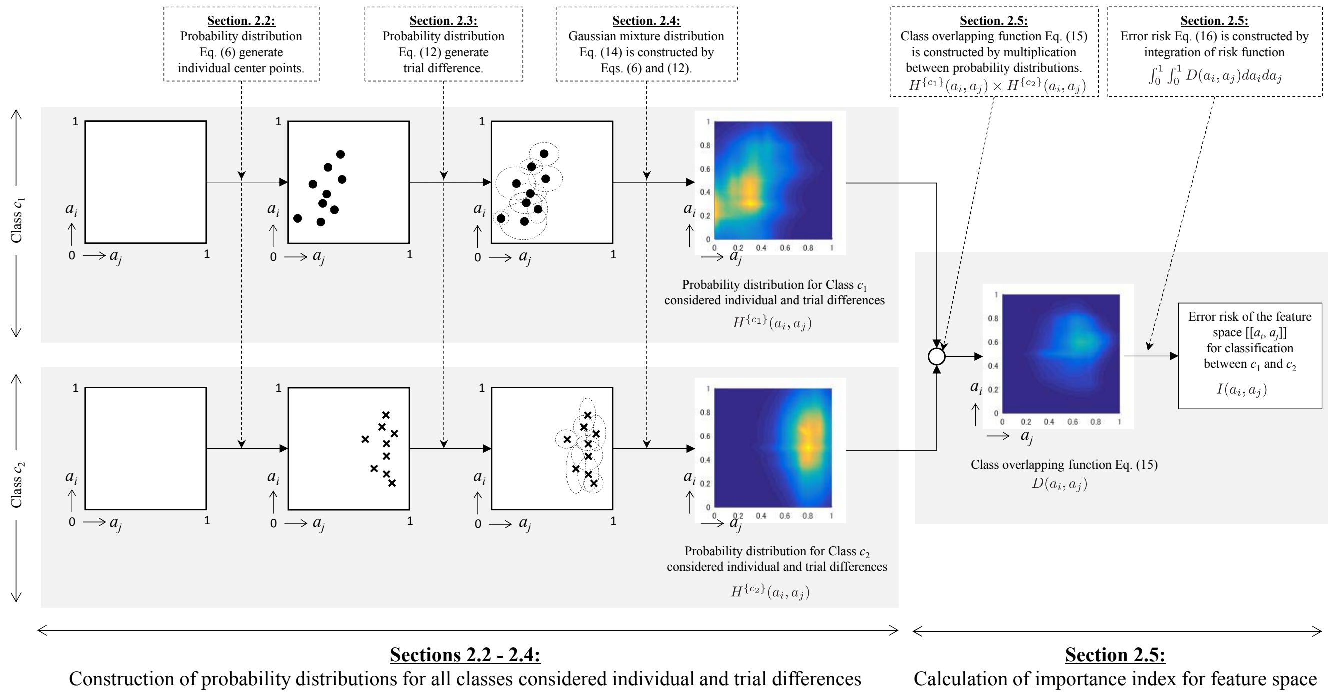

Section. 2.2:
Probability distribution Eq. (6) generate individual center points.
Section. 2.3:
Probability distribution Eq. (12) generate trial difference.
Section. 2.4:
Gaussian mixture distribution Eq. (14) is constructed by Eqs. (6) and (12).
Section. 2.5:
Class overlapping function Eq. (15) is constructed by multiplication between probability distributions.
$H^{\{c_1\}}(a_i,a_j) \times H^{\{c_2\}}(a_i,a_j)$
Section. 2.5:
Error risk Eq. (16) is constructed by integration of risk function
$\int_0^1\int_0^1 D(a_i,a_j)da_ida_j$
Class c_1
Class c_2
a_i
a_j
0
1
0.2
0.4
0.6
0.8
Probability distribution for Class c_1 considered individual and trial differences
$H^{\{c_1\}}(a_i,a_j)$
Probability distribution for Class c_2 considered individual and trial differences
$H^{\{c_2\}}(a_i,a_j)$
Class overlapping function Eq. (15)
$D(a_i,a_j)$
Error risk of the feature space $[[a_i, a_j]]$ for classification between c_1 and c_2
$I(a_i,a_j)$
Sections 2.2 - 2.4:
Construction of probability distributions for all classes considered individual and trial differences
Section 2.5:
Calculation of importance index for feature space

Fig. 2. Principles of importance indices for feature spaces with introduced individual and trial differences.

dinates, which represent individual differences, represents trial differences.) Using the actually obtained data for M subjects, we construct probability distributions to spread a single subject's barycentric coordinates in the feature space. The probability density function is a multivariate normal distribution type, the same as in the preceding Section 2.2.

The variances of the features a_i and a_j of subject m in Class c_n are denoted by $\sigma_{a_i,m}^{\{c_n\}}$, $\sigma_{a_j,m}^{\{c_n\}}$. They represent variances in the plural trials for subject m of Class c_n motions in the feature space $[\![a_i, a_j]\!]$, and correspond to trial differences. Then, the mean value vector $\boldsymbol{u}_{a_{i,j}}^{\{c_n\}}$ for all subject trial differences in the feature space $[\![a_i, a_j]\!]$ and variance-covariance matrix $S_{a_{i,j}}^{\{c_n\}}$ are expressed by the following equations:

$$\boldsymbol{u}_{a_{i,j}}^{\{c_n\}} = \left(u_{a_i}^{\{c_n\}}\ u_{a_j}^{\{c_n\}}\right)^T, \quad \dots\dots\dots\dots (7)$$

$$S_{a_{i,j}}^{\{c_n\}} = \begin{pmatrix} s_{a_i}^{\{c_n\}\ 2} & s_{a_{i,j}}^{\{c_n\}} \\ s_{a_{i,j}}^{\{c_n\}} & s_{a_j}^{\{c_n\}\ 2} \end{pmatrix}, \quad \dots\dots\dots (8)$$

where $u_{a_i}^{\{c_n\}}$, $u_{a_j}^{\{c_n\}}$, $s_{a_i}^{\{c_n\}\ 2}$ and $s_{a_{i,j}}^{\{c_n\}}$ respectively, denote the mean value, variance, and covariance for all subjects in relation to each subject's variance in Class c_n in the feature space $[\![a_i, a_j]\!]$. These are defined by the following equations:

$$u_{a_i}^{\{c_n\}} = \frac{1}{M}\sum_{m=1}^{M} \sigma_{a_i,m}^{\{c_n\}}, \quad \dots\dots\dots\dots (9)$$

$$s_{a_i}^{\{c_n\}\ 2} = \frac{1}{M}\sum_{m=1}^{M} (\sigma_{a_i,m}^{\{c_n\}} - u_{a_i}^{\{c_n\}})^2, \quad \dots\dots (10)$$

$$s_{a_{i,j}}^{\{c_n\}} = \frac{1}{M}\sum_{m=1}^{M} (\sigma_{a_i,m}^{\{c_n\}} - u_{a_i}^{\{c_n\}})(\sigma_{a_j,m}^{\{c_n\}} - u_{a_j}^{\{c_n\}}). \quad .. (11)$$

With the mean value vector $\boldsymbol{u}_{a_{i,j}}^{\{c_n\}}$ and variance-covariance matrix $S_{a_{i,j}}^{\{c_n\}}$ as parameters of the multivariate normal distribution, for an unknown subject m' trial difference $\boldsymbol{x}_{m'}$ in the feature space $[\![a_i, a_j]\!]$ is generated by following the probability distribution described below:

$$\begin{aligned} F_{\text{Trial}}^{\{c_n\}}(\boldsymbol{x}_{m'}) = {} & \frac{1}{(\sqrt{2\pi})^2\sqrt{|S_{a_{i,j}}^{\{c_n\}}|}} \times \\ & \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_{m'} - \boldsymbol{u}_{a_{i,j}}^{\{c_n\}}\right)^T \times \right. \\ & \left. S_{a_{i,j}}^{\{c_n\}\ -1}\left(\boldsymbol{x}_{m'} - \boldsymbol{u}_{a_{i,j}}^{\{c_n\}}\right)\right). \quad \dots (12) \end{aligned}$$

$\boldsymbol{x}_{m'}$ is a two-dimensional vector with the variance of features a_i, a_j in plural trials of subject m' of Class c_n motions as a component. The third feature space from the left of **Fig. 2** shows that we have given the variance generated by following the probability distribution $F_{\text{Trial}}^{\{c_n\}}(\boldsymbol{x}_{m'})$ we have constructed in this section to the individual barycentric coordinates generated by following the probability distribution $F_{\text{Trial}}^{\{c_n\}}(\boldsymbol{x}_{m'})$.

We have generated variances by following probability distributions, so that such variances sometimes have negative values. If the probability distributions $F_{\text{Trial}}^{\{c_n\}}(\boldsymbol{x}_{m'})$ used to generate individual trial differences should generate negative variances, they would have the value of 0.

2.4. Probability Distributions of Class c_n Through Introduction of Individual/Trial Differences

Equation (6) described in Section 2.2 enables us to generate the barycentric coordinates $\boldsymbol{a}_{i,j,m'}$ of subject m' in the feature space $[\![a_i,a_j]\!]$. Eq. (12) described in Section 2.3 further enables us to introduce trial differences $\boldsymbol{x}_{m'}$ into the barycentric coordinates of generated subject m'.

Then, features $\boldsymbol{a}_{m'} = (a_i\ a_j)^T$ to be calculated for the plural trials of subject m' of Class c_n motions are generated in the feature space $[\![a_i,a_j]\!]$ by the following multivariate normal distributions, which follow parameters $\boldsymbol{a}_{i,j,m'}$ and $\boldsymbol{x}_{m'}$:

$$G^{\{c_n\}}(\boldsymbol{a}_{m'}) = \frac{1}{(\sqrt{2\pi})^2\sqrt{|\boldsymbol{x}_{m'}|}} \times \exp\Big(-\frac{1}{2}(\boldsymbol{a}_{m'}-\boldsymbol{a}_{i,j,m'})^T \times \boldsymbol{x}_{m'}^{-1}(\boldsymbol{a}_{m'}-\boldsymbol{a}_{i,j,m'})\Big). \quad \ldots\ldots (13)$$

Thus, the probability distribution enables us to generate the features for the plural trials of subject m' of Class c_n motions.

The features of Class c_n motions of M' subjects are plotted in the feature space $[\![a_i,a_j]\!]$ by the following probability distribution:

$$H^{\{c_n\}}(a_i,a_j) = \frac{1}{M'}\sum_{m'=1}^{M'} G^{\{c_n\}}(\boldsymbol{a}_{m'}). \quad \ldots\ldots (14)$$

$G^{\{c_n\}}(\boldsymbol{a}_{m'})$ is a multivariate normal distribution and $H^{\{c_n\}}(a_i,a_j)$, standardized for the total sum to be a probability distribution, should have a mixed Gaussian distribution.

The fourth feature space from the left of **Fig. 2** depicts the probability distribution $H^{\{c_n\}}(a_i,a_j)$ of Class c_n, in which individual and trial differences are introduced.

2.5. Class Overlapping Function and Error Risk

We introduce evaluation indices for feature spaces, using the probability distribution $H^{\{c_n\}}(a_i,a_j)$ of Class c_n, where individual and trial differences are introduced. This corresponds to the "calculation of importance index for feature space" described on the right side of **Fig. 2**.

The feature space $[\![a_i,a_j]\!]$, where a comparison of the probability distributions of Classes c_1 and c_2 (with individual and trial differences introduced therein) shows the coordinates of a class with a high occurrence probability of overlapping one another, may be regarded as a feature space where misclassifications are more likely to be caused. We have worked out the following function to represent the likelihood of the coordinates in the feature space causing misclassifications:

$$D(a_i,a_j) = H^{\{c_1\}}(a_i,a_j) \times H^{\{c_2\}}(a_i,a_j). \quad \ldots\ldots (15)$$

Any $H^{\{c_n\}}(a_i,a_j)$, being a probability density function, has positive values over the entire domain. Hence, function $D(a_i,a_j)$, obtained by multiplying the probability density function, is also assured to be positive in value over the entire domain. If the coordinates of Class c_1 or c_2 with a higher occurrence probability overlap one another, $D(a_i,a_j)$ will have a higher value. Therefore, we refer to $D(a_i,a_j)$ as a class overlapping function.

We define a function integrating $D(a_i,a_j)$ over the entire domain. This represents the error risk of features a_i and a_j in the class separation problem for Classes c_1 and c_2:

$$I(a_i,a_j) = \int_0^1\int_0^1 D(a_i,a_j)da_ida_j. \quad \ldots\ldots (16)$$

Error risk $I(a_i,a_j)$ corresponds to the volume of the class overlapping function $D(a_i,a_j)$. Being a one-dimensional real-valued function, it has a higher value when the coordinates of classes with higher occurrence probabilities overlap one another. In other words, a feature space with the smallest error risk is meant to be the most important feature space for the separation of Classes c_1 and c_2. Incidentally, Eq. (16) can be approximated by the following equation:

$$I(a_i,a_j) \simeq \sum_{a_i\in[0,1]}\sum_{a_j\in[0,1]} D(a_i,a_j). \quad \ldots\ldots (17)$$

As a feature space with a smaller error risk $I(a_i,a_j)$, the coordinates of a class with a higher occurrence probability do not overlap. Two optimal features a^{opt1} and a^{opt2} for separating Classes c_1 and c_2 can be obtained by solving the following optimization problem:

$$[\![a^{\mathrm{opt1}},a^{\mathrm{opt2}}]\!] = \operatorname*{argmin}_{a_i,a_j} I(a_i,a_j). \quad \ldots\ldots (18)$$

The heat map on the left side of **Fig. 3** exhibits the class overlapping function $D(a_i,a_j)$ of feature space $[\![a_i,a_j]\!]$, composed of features a_i and a_j. Deeper blue colors (darker region in the case of gray scale print) indicate lower values and deeper yellow colors (lighter region) indicate higher values.

$D(a_1,a_2)$ has generally low values across the entire feature space. This indicates either that the coordinates of Classes c_1 and c_2, which have high occurrence probabilities, overlap one another almost nowhere on the entire feature space, or that the feature space has a low error risk. Therefore, the error risk $I(a_1,a_2)$, which is obtained from the definite integral over the entire domain, should be low.

$D(a_1,a_3)$ has generally high values across the entire feature space. This indicates either that the coordinates of Classes c_1 and c_2, which have high occurrence probabilities, overlap one another nearly on the entire feature space, or that the feature space has a high error risk. Therefore, the error risk $I(a_1,a_3)$ obtained from the definite integral over the entire domain should be high.

$D(a_1,a_4)$ has generally high values in regions where a_1 is low and a_4 is high. This indicates that the coordinates of Classes c_1 and c_2 with high occurrence probabilities overlap one another at the lower right. With the coordinates with high occurrence probabilities overlapping one another, the error risk $I(a_1,a_4)$ obtained from the integral of $D(a_1,a_4)$ over the entire domain should be high.

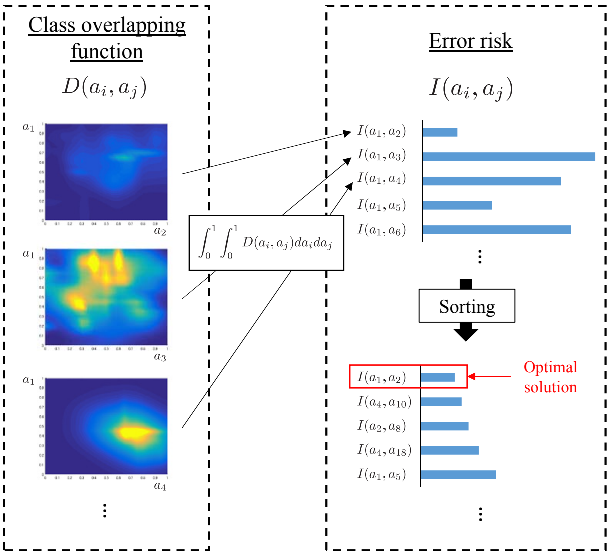


Fig. 3. Relationships between Class overlapping function, error risk, and optimal solution.

Apply the operations to all combinations of selection candidate features and, in ascending order, sort them as $I(a_i,a_j)$, so that you can place the feature spaces in the order of error risks. As the optimal solution in **Fig. 3** shows, the combination of features at the top creates a feature space with the lowest error risk.

2.6. Multiclass Separation Problem on Multidimensional Feature Space

We generalize our proposed method. Specifically, we perform the following function: "In a $C_{\max}$ class separation problem with Classes $c_1,\ldots,c_{C_{\max}}$, compose an optimal ν-dimensional feature space $[\![a^{\text{opt}1},a^{\text{opt}2},\ldots,a^{\text{opt}\nu}]\!]$ out of a set $A=\{a_1,\ldots,a_p\}$ with p features as members." p and ν are natural numbers and $\nu<p$.

To search an optimal feature space, we calculate the indices for evaluating the ν-dimensional feature space $[\![a_{i_1},a_{i_2},\ldots,a_{i_\nu}]\!]$ $(i_1,i_2,\ldots,i_\nu=1,\ldots,p, i_1\neq i_2,\ldots,\neq i_\nu)$, which is composed of $a_{i_1},a_{i_2},\ldots,a_{i_\nu}\in A$ in solving the class separation problem. Additionally, we select a feature space that has the best index as an optimal feature space $[\![a^{\text{opt}1},a^{\text{opt}2},\ldots,a^{\text{opt}\nu}]\!]$, $(a^{\text{opt}1},a^{\text{opt}2},\ldots,a^{\text{opt}\nu}\in A)$.

As in Sections 2.2 to 2.4, we pick ν features $a_{i_1},a_{i_2},\ldots,a_{i_\nu}$ $(i_1,i_2,\ldots,i_\nu=1,\ldots,p,i_1\neq i_2,\ldots,\neq i_\nu)$ out of a set $A=\{a_1,\ldots,a_p\}$ with p features as members, and calculate a mixed Gaussian distribution $H^{\{c_n\}}(a_{i_1},a_{i_2},\ldots,a_{i_\nu})$ of c_n $(n=1,\ldots,C_{\max})$ in the ν-dimensional feature space $[\![a_{i_1},\ldots,a_{i_\nu}]\!]$. Then, we can grasp the occurrence probabilities of the respective classes on feature space $[\![a_{i_1},a_{i_2},\ldots,a_{i_\nu}]\!]$.

The overlapping of coordinates with high occurrence probabilities in each class indicates higher error risks. Therefore, we can obtain a function for expressing which regions of the feature space are likely to misclassify how much by multiplying the mixed Gaussian distributions $H^{\{c_n\}}(a_{i_1},a_{i_2},\ldots,a_{i_\nu})$:

$$D(a_{i_1},a_{i_2},\ldots,a_{i_\nu})=\prod_{n=1}^{C_{\max}} H^{\{c_n\}}(a_{i_1},a_{i_2},\ldots,a_{i_\nu}). \quad (19)$$

This is the class overlapping function of feature space $[\![a_{i_1},a_{i_2},\ldots,a_{i_\nu}]\!]$ composed of ν features $a_{i_1},a_{i_2},\ldots,a_{i_\nu}$ $(i_1,i_2,\ldots,i_\nu=1,\ldots,p,i_1\neq i_2,\ldots,\neq i_\nu)$ picked from a set $A=\{a_1,\ldots,a_p\}$ with p features as members.

We calculate the error risks $I(a_{i_1},a_{i_2},\ldots,a_{i_\nu})$ representing the overall error risk of feature space $[\![a_{i_1},a_{i_2},\ldots,a_{i_\nu}]\!]$. This is defined by the ν-multiple integral of the class overlapping function, calculated from the combinations of all features. Because a feature space with the smallest error risk is the most desirable, the optimal feature space can be obtained from the following equation:

$$I(a_{i_1},a_{i_2},\ldots,a_{i_\nu})=\underbrace{\int_0^1\cdots\int_0^1}_{\nu} D(a_{i_1},a_{i_2},\ldots,a_{i_\nu})da_{i_1}\ldots da_{i_\nu}, \quad (20)$$

$$[\![a^{\text{opt}1},\ldots,a^{\text{opt}\nu}]\!]=\underset{a_{i_1},a_{i_2},\ldots,a_{i_\nu}}{\operatorname{argmin}}\; I(a_{i_1},a_{i_2},\ldots,a_{i_\nu}).(21)$$

The calculation frequency number for calculating error risks $I(a_{i_1},a_{i_2},\ldots,a_{i_\nu})$ to search ν effective features out of p features is ${}_{C_{\max}}\mathrm{C}_\nu$. We should note, therefore, that the calculation time will become exponentially long if the scale of the problem is too large.

3. Evaluation Experiments

3.1. Experimental Purpose and Outline

We verify whether the indices for evaluating the feature spaces proposed in this paper can really select better features than existing methods. The problem we have assigned in the evaluation experiment is classifying butterfly and breaststroke from one another using the inertia sensor data acquired during a swimming race. The swimming style classifier constructed in many earlier studies have commonly confused butterfly and breaststroke [17–20]. This is attributable to the fact that butterfly and breaststroke are more similar than other swimming motions. This was adopted because a problem with higher solution difficulty is better suited for evaluating the proposed method.

We apply the proposed method to solve the problem of converting inertia sensor data into various features, and selecting from among them two features effective at classifying between butterfly and breaststroke.

To evaluate how our proposed method is an improved solution, we have solved the same problem using existing feature space evaluation indices such as BW-Ratio [7], OOB [8], ReliefF [9], and Minimum Reference Set (MRS) [10]. Among the existing feature space evaluation indices, BW-Ratio, ReliefF, and OOB are widely used as general methods for selecting features, and are believed to

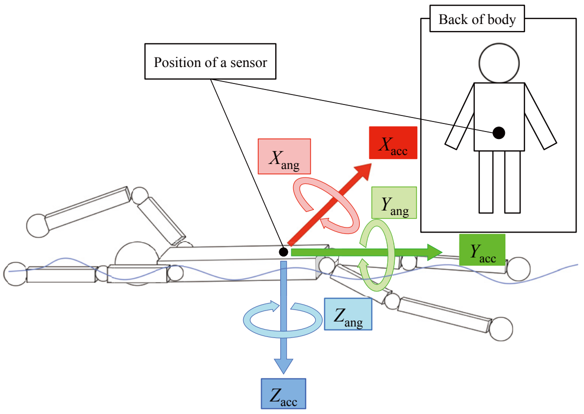


Fig. 4. Sensor mounting position and axial settings.

Table 1. Adopted features $a_i (i = 1, \ldots, 42)$.

Definition	Acceleration		Angular velocity	
	i	ζ	i	ζ
Mean $a_i = \mathrm{Mean}(\zeta)$	1	X_{acc}	22	X_{ang}
	2	Y_{acc}	23	Y_{ang}
	3	Z_{acc}	24	Z_{ang}
Variance $a_i = \mathrm{Var}(\zeta)$	4	X_{acc}	25	X_{ang}
	5	Y_{acc}	26	Y_{ang}
	6	Z_{acc}	27	Z_{ang}
Skew $a_i = \mathrm{Skew}(\zeta)$	7	X_{acc}	28	X_{ang}
	8	Y_{acc}	29	Y_{ang}
	9	Z_{acc}	30	Z_{ang}
Kurtosis $a_i = \mathrm{Kurt}(\zeta)$	10	X_{acc}	31	X_{ang}
	11	Y_{acc}	32	Y_{ang}
	12	Z_{acc}	33	Z_{ang}
Maximum $a_i = \mathrm{Max}(\zeta)$	13	X_{acc}	34	X_{ang}
	14	Y_{acc}	35	Y_{ang}
	15	Z_{acc}	36	Z_{ang}
Minimum $a_i = \mathrm{Min}(\zeta)$	16	X_{acc}	37	X_{ang}
	17	Y_{acc}	38	Y_{ang}
	18	Z_{acc}	39	Z_{ang}
Frequency Domain Entropy $a_i = \mathrm{Ent}(\zeta)$	19	X_{acc}	40	X_{ang}
	20	Y_{acc}	41	Y_{ang}
	21	Z_{acc}	42	Z_{ang}

M_{acc} : M-axial acceleration
M_{ang} : M-axial angular velocity

be highly reliable (for instance, BW-Ratio is used in References [11, 12], OOB, in References [13, 14], and ReliefF, in References [15, 16]). In contrast, MRS selects features using a small amount of training data, which is one of the aims of this paper. We followed the following procedure: extract some data out from the training data and check whether such extracted data can classify all training data using the 1-nearest neighbor method; then, any feature space where all training data can be classified without errors with as little data as possible is interpreted as a better space.

To evaluate the methods, we conducted experiments using a total of 23 student subjects (nine males and four females) in a swimming club of the university. The subjects are 19.9±1.7 years, 168.8±6.6 cm tall, weigh 63.4±5.3 kg, and have 14.2±3.9 years of swimming experience. The pool we used for the experiments is a short course (25 m). In conducting the experiments, we told the subjects that the "acquired data are used for research purposes only" and conducted the experiments only with swimming club members who consented to the research.

In the experiments, we used sensors made by Sports Sensing Co., Ltd. [24]. The specifications of the sensors are: acceleration (±5 G); angular velocity (±1500 dps); acquisition of terrestrial magnetism information; sampling frequency 100 Hz; mass 20 g; size 67 mm × 26 mm × 8 mm. Acquired data are stored in built-in memory (32 MB). For more detailed specifications, refer to the product catalog (Waterproof 9-Axial Wireless Motion Sensor (5 G/1500 dps), Model SS-WS1215, Type A).

Figure 4 shows the position of the sensor and axial settings, where X_{acc} denotes the acceleration of the X-axis and X_{ang} denotes the angular velocity of the X-axis; the same notation applies to the Y-axis and Z-axis. In conducting the experiments, the subjects were asked to select two swimming styles they are good at from four swimming styles, and the subjects who selected either butterfly or breaststroke performed the motions. They were instructed to lap swim a 25 m pool (50 m in total) with full force. We taped the swimming motions with a video camera (30 fps) in order to collate the sensor data waveforms and swimming motions. We used a Sony digital HD video camera recorder HDR-CX720V [25].

3.2. Results and Discussion

From the experiments conducted using the conditions, we acquired data from four subjects for breaststroke and six subjects for butterfly.

3.2.1. Conversion into Features

We converted inertia sensor acquired data into features using the sliding window method [21, 22]. In constructing a classifier to classify swimming styles, the window width was decided from the time required for one stroke during a swimming race [23]: construct normal distributions of stroke time for all swimming styles and calculate their total sum, to obtain a stroke time with the highest occurrence probability that can be applied to any swimming style. The selected window width was 106 sample points and the selected slide width was 53 sample points (half of the window width).

Convert sensor data in the said window width into features. We have used 42 types of features ($A = \{a_1, \ldots, a_{42}\}$), as shown in **Table 1**. These are the same features used in earlier studies on swimming style classification [17–20].

We consecutively converted sensor data into features using the procedures, and obtained 162 and 210 feature vector points for breaststroke and butterfly, respectively. In the proposed method, in view of the domain for defining the multiple integral, the range of each feature must be standardized to $[0,1]$. Therefore, the maximum and minimum values for each feature were searched. The maximum value was standardized as 1 and the minimum value was standardized as 0.

3.2.2. Separation of Training and Test Data

We separated the acquired data for four subjects for breaststroke and six subjects for butterfly into training and test data. Training data were used to select features and construct a classifier. Test data were used to verify the generalization performance of the proposed method. For breaststroke, the data for three subjects were separated as training data, and the data from one subject were used as test data. For butterfly, the data from five subjects were separated as training data and the data from one subject were used as test data. To conduct the evaluations under the strictest conditions possible, we selected one subject each for breaststroke and butterfly, whose feature vectors are most different from those of the other subjects. The following is the process for selecting data from the single subject (for both breaststroke and butterfly) as the test data.

With "br" denoting breaststroke and "bu" denoting butterfly; $a_i(\gamma_{\beta^k})$ value means γ_{β^k}-th features of subject β^k in swimming $k \in \{\text{br}, \text{bu}\}$. The total features obtained by applying the sliding window method to the sensor data from subject β^k means Γ_{β^k}. Then, the mean value $\overline{a_i(\beta^k)}$ of the features a_i of subject β^k in swimming style k is expressed as follows:

$$\overline{a_i(\beta^k)} = \frac{1}{\Gamma_{\beta^k}} \sum_{\gamma_{\beta^k}=1}^{\Gamma_{\beta^k}} a_i(\gamma_{\beta^k}). \quad \cdots\cdots\cdots \quad (22)$$

Calculate the mean value for each swimming style of for all the subjects. Then, calculate the mean value $\overline{a_i^k}$ of for all the subjects' features a_i in for swimming style k:

$$\overline{a_i^k} = \frac{1}{M_k} \sum_{\beta^k=1}^{M_k} \overline{a_i(\beta^k)}. \quad \cdots\cdots\cdots\cdots \quad (23)$$

M_k denotes the number of subjects for swimming style $k \in \{\text{br}, \text{bu}\}$, where $M_{\text{br}} = 4$ and $M_{\text{bu}} = 6$. Term $\overline{a_i^k}$ is the mean value of the features a_i of all subjects for swimming style k, and is a representative value of features a_i for swimming style k. By subtracting from $\overline{a_i^k}$ the mean value $\overline{a_i(\beta^k)}$ of features a_i for subject β_k, we can quantitatively express the extent to which subject β_k deviates from the representative value of features a_i. Therefore, we define the deviations in features a_i of subject β_k in swimming style k as follows:

$$\Delta a_i(\beta^k) = |\overline{a_i^k} - \overline{a_i(\beta^k)}| \quad \cdots\cdots\cdots \quad (24)$$

We calculate such deviations in for all of the features $a_1, \ldots, a_{42}$ and calculate their total $\Delta A(\beta^k)$ as follows:

$$\Delta A(\beta^k) = \sum_{i=1}^{42} \Delta a_i(\beta^k). \quad \cdots\cdots\cdots\cdots \quad (25)$$

$\Delta A(\beta^k)$ is obtained by totaling the deviations of features a_i of subject β_k, and can represent the extent to which subject β_k deviates from the other subjects.

We calculated $\Delta A(\beta^k)$ for all subjects for swimming style k and selected the subject data with the largest $\Delta A(\beta^k)$ as the test data. We separated the data for the four breaststroke subjects into three training data and one test data and the data for the six butterfly stroke subjects into five training data and one test data. Because the test data represent the data for a subject whose features most deviate in value from the others, the separation problem we assigned should be relatively more difficult than that for randomly selected test data.

3.2.3. Feature Selection Results and Classification Accuracy

We discuss the problem of selecting two features (from 42 total) that effectively classify between breaststroke and butterfly. There are ${}_{42}C_2 = 861$ solution patterns. We selected two features important for discriminating butterfly from breaststroke, by calculating the feature space evaluation indices for all of the solution patterns using both the proposed method and existing methods (BW-Ratio, OOB, and ReliefF). One of the parameters we used in applying the proposed method was the number of people M', which was generated following the probability distribution $G^{\{c_n\}}(\boldsymbol{a}_{m'})$. In view of computational complexity, we adopted $M' = 100$. The feature selection results are listed in the first row of **Table 2**. Because the importance of feature combinations is available for BW-Ratio and the proposed method, the top five important combinations of features are listed. In OOB and ReliefF, where importance per feature is output, we selected the first and second important features for the first important feature space, the third and fourth important features for the second important feature space, and so on for the top five feature spaces. We can see from **Table 2** that features a_3, a_7 and a_8 are regarded as important in any method. Because any method is designed to improve generalization performance, some similar features are reasonably selected using any method.

We constructed a classifier to classify between the butterfly and breaststroke using SVM on training data for the selected feature spaces. We used the following three kernel functions: linear kernel (linear), Gaussian kernel (RBF), and poly-nominal kernel (poly). We set two coefficients to impose a penalty on training data misclassifications: $c = 1$ and $c = 100$. $c = 1$ corresponds to a soft-margin SVM that imposes little penalty on training data misclassifications and $c = 100$ corresponds to a hard-margin SVM that imposes a great penalty on training data misclassifications. The hard margin SVM, which is deeply matched to training data, can generate complex classification criteria, but is more susceptible to overlearning. In contrast, the soft margin SVM, which is not overmatched to training data, is less susceptible to overlearning, and is more likely to improve generalization performance.

We constructed six types of SVM using a combination of three types of kernel functions and two types of mar-

Table 2. F-measures of test data for respective classifiers on feature spaces composed of top 5 indices.

Feature		Selected	Classification condition (SVM)									
selection	Top 5	features	Hard margin ($c = 100$)			Soft margin ($c = 1$)			Mean 1	Mean 2	Max	Min
algorithm			Linear	RBF	Poly	Linear	RBF	Poly				
BW-Ratio [7]	1	$\llbracket a_7, a_{38} \rrbracket$	.902	.738	.727	.885	.780	.780	.802	.787	.957	.281
	2	$\llbracket a_7, a_8 \rrbracket$	.957	.936	.957	.957	.957	.957	.953			
	3	$\llbracket a_8, a_{37} \rrbracket$	.957	.281	.339	.957	.550	.368	.575			
	4	$\llbracket a_7, a_{31} \rrbracket$	.894	.375	.375	.894	.773	.773	.680			
	5	$\llbracket a_7, a_{17} \rrbracket$	.936	.936	.898	.913	.936	.917	.923			
OOB [8]	1	$\llbracket a_8, a_{24} \rrbracket$	.080	.000	.000	.154	.080	.080	.066	.345	.723	.000
	2	$\llbracket a_{27}, a_{34} \rrbracket$	.000	.000	.000	.000	.000	.000	.000			
	3	$\llbracket a_7, a_{39} \rrbracket$	.550	.723	.723	.636	.571	.667	.645			
	4	$\llbracket a_3, a_{38} \rrbracket$	.571	.524	.524	.558	.558	.558	.549			
	5	$\llbracket a_{15}, a_{16} \rrbracket$	.450	.444	.444	.488	.488	.474	.465			
ReliefF [9]	1	$\llbracket a_8, a_{34} \rrbracket$	.452	.452	.452	.452	.452	.452	.452	.654	.980	.347
	2	$\llbracket a_3, a_7 \rrbracket$	.957	.762	.716	.957	.885	.885	.860			
	3	$\llbracket a_2, a_{18} \rrbracket$	.960	.909	.958	.960	.980	.980	.958			
	4	$\llbracket a_{19}, a_{28} \rrbracket$	.578	.667	.649	.571	.585	.585	.606			
	5	$\llbracket a_{12}, a_{20} \rrbracket$	.410	.390	.368	.450	.390	.347	.392			
MRS [10]	1	$\llbracket a_3, a_{37} \rrbracket$	.558	.558	.558	.558	.558	.558	.558	.535	1.000	.276
	2	$\llbracket a_{36}, a_{37} \rrbracket$	.400	.276	.276	.400	.276	.387	.336			
	3	$\llbracket a_8, a_{22} \rrbracket$	1.000	.588	.535	1.000	.844	.760	.788			
	4	$\llbracket a_{15}, a_{34} \rrbracket$	.452	.452	.452	.452	.452	.452	.452			
	5	$\llbracket a_{14}, a_{28} \rrbracket$	.571	.529	.571	.571	.485	.529	.543			
Ours	1	$\llbracket a_3, a_7 \rrbracket$	.957	.762	.716	.957	.885	.885	.860	.873	.980	.558
	2	$\llbracket a_3, a_8 \rrbracket$	.980	.941	.941	.980	.960	.941	.957			
	3	$\llbracket a_7, a_8 \rrbracket$	.957	.936	.957	.957	.957	.957	.953			
	4	$\llbracket a_3, a_{18} \rrbracket$	.980	.960	.941	.980	.960	.960	.963			
	5	$\llbracket a_3, a_5 \rrbracket$	.558	.787	.774	.558	.558	.558	.632			

Linear: Linear kernel SVM, RBF: Gaussian kernel SVM, Poly: Polynomial kernel SVM

gins for feature spaces composed of the top five features extracted by each method. Additionally, we calculated the F-measures of the test data. The results are listed in the 4th to 13th rows of **Table 2**. The Mean 1 values are the mean F-measures of six types of classifiers for the individual feature spaces. Mean 2 is the mean of the six Mean 1 values, and represents the overall generalization performance of the feature spaces obtained from the respective methods. Max and Min indicate the maximum and minimum F-measures for all feature classifiers obtained from the respective methods. A comparison of the Mean 1 values for the respective methods shows that the proposed method is not necessarily more accurate than other methods, but a comparison of the Mean 2 values shows that the proposed method is the most accurate. In terms of the maximum F-measures of a total of 30 types of classifiers (a combination of top five feature spaces and six patterns), MRS, an existing method, has the highest value, and the proposed method has the highest value in terms of minimum F-measures. In terms of the F-measures for optimal solutions by the various methods, the proposed method also produces the best result. The highest mean value (Mean 2) means that the method has the highest expectation for generalization performance; the highest minimum value means that it is less likely to obtain feature spaces with low generalization performance; the highest F-measure for the optimal solutions means that the best solution is obtained by selecting just one pair of features and automatically applying the methods.

Table 3. Comparison of mean F-measures obtained from 30 types of classifiers.

Comparison			def	p-value
Ours	vs	BW-Ratio	+.087	*
Ours	vs	ReliefF	+.220	**
Ours	vs	OOB	+.528	**
Ours	vs	MRS	+.338	**

*:$p < .05$, **:$p < .01$

Incidentally, OOB produced extremely low values. OOB is a method for calculating the importance of features using increases in errors by randomly varying the values of features. This is done to calculate the importance of the pseudo test data generated by the bootstrap sampling method. In this paper, we have selected as test data from a subject whose feature vector most deviated from those of other subjects. It is anticipated, therefore, that the pseudo test data by OOB and the real test data are different from one another in the feature space, and that the importance of features as measured on pseudo test data is not well reflected in the accuracy of real test data. To conduct a more quantitative evaluation, we have compared the Mean 2 t-test values for the proposed method and each existing method. **Table 3** shows the comparison

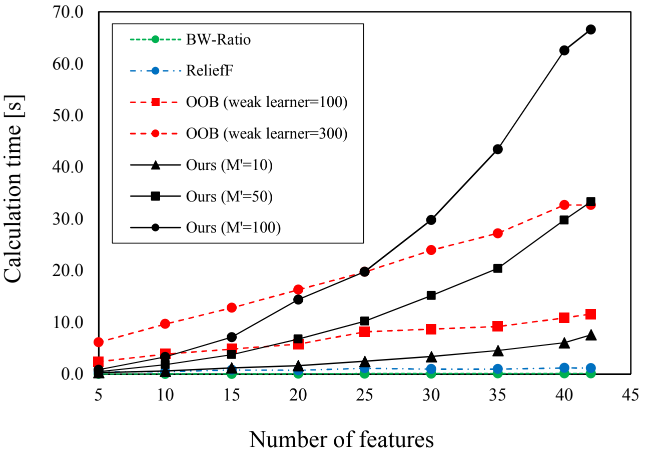


Fig. 5. Calculation time by respective methods (Proposed Method, BW-Ratio, OOB, ReliefF).

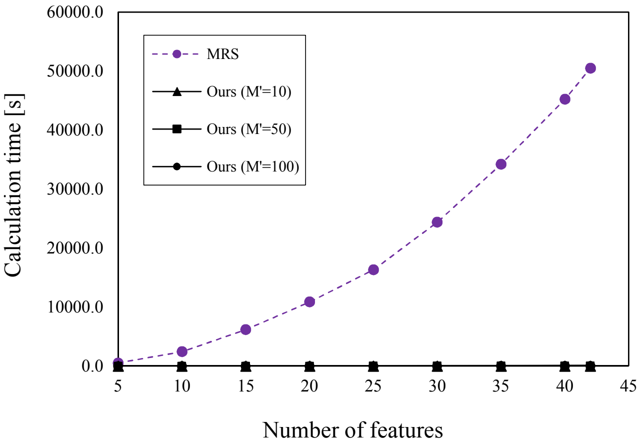


Fig. 6. Calculation time by respective methods (Proposed Method, MRS).

results. The second row of **Table 3** shows the subtractions of the Mean 2 values of the existing methods from those of the proposed method. A comparison between the proposed method and BW-Ratio method shows that the proposed method has a significant difference (at a 5% level) in F-measures in comparison with OOB, ReliefF and MRS shows the proposed method have a significant difference of 1% level. The experimental results suggest that the proposed feature space evaluation indices have the potential to select features with higher generalization performance than existing methods, in a context in which training data can only be acquired from a small number of subjects.

3.3. Calculation Time

We evaluate the calculation time required by the proposed method. We have executed the evaluations, using a the same computer for both the proposed method and the existing methods in the following environment: OS: Windows 8.1, CPU: Intel(R) Core(TM) i7-4510 CPU (2.00 GHz-2.59 GHz), RAM 8.00 GB. The data we used were the training data described in Section 3.2.2. We measured the calculation time after converting the sensor data into features, so that conversion time was not included in the compuation time.

The evaluation results are shown in **Figs. 5** and **6**. **Fig. 5** shows a comparison of calculation times between the proposed method, BW-Ratio, OOB, and ReliefF. **Fig. 6** shows a comparison between the proposed method and MRS.

With the searched space size indicated on the axis of abscissa and the calculation time on the axis of ordinate, **Figs. 5** and **6** show the lengths of time taken to select two effective features from the overall feature set (5 to 42) indicated on the axis of abscissa. The different colors and lines indicate different methods: green dashed line with circle for BW-Ratio, blue dashed dot line with circle for ReliefF, red dashed line with circle or square for OOB, purple dashed with circle in **Fig. 6** for MRS, and black solid line with circle, triangle or square for the proposed method. The forms of the markers indicate the heuristic parameters of the different methods: red dashed line with square indicates that 100 weak learning machines are constructed by OOB and red dashed line with circle indicates that 300 weak learning machines are constructed. The more weak learning machines are constructed, the fewer decision surfaces are generated to separate the noise data present in training data as a result of decreased variance. This tends to improve generalization performance. On the other hand, as the number of weak learning machines increases, more decision trees must be constructed, which directly leads to increased calculation time.

The individual barycentric coordinates M' generated by the probability distributions in the proposed method are indicated by a black solid line with circle (100 persons), black line with square (50 persons), or black solid line with triangle (10 persons). As the individual barycentric coordinates are generated for more individuals, more mixed Gaussian distributions $H_{c_n}(a_i, a_j)$ with a high capacity for representing unknown data are generated, which will increase the possibility of discovering more appropriate features. On the other hand, the more persons, the more probability distributions must be constructed, which will increase calculation time.

Considering these factors, we interpret the obtained results. We can see from **Fig. 5** that BW-Ratio and ReliefF take an extremely short time (approximately one second) to select two effective features, despite the large number of features that must be considered. In contrast, OOB and the proposed method require a longer calculation time than BW-Ratio. OOB requires a longer calculation time because it must construct classifiers by ensemble learning the feature spaces for all features. To measure the importance of features, we constructed classifiers using a random forest consisting of a number of decision trees. In constructing decision trees, OOB must solve the optimization problem to search for features that can minimize information entropy. OOB has solved said optimization problem an enormous number of times, which appears to have required a longer calculation time than ReliefF or BW-Ratio. The proposed method seems to have taken a long time to construct a large number of probability distributions. We can see, however, that the proposed method and OOB can adjust their calculation times using heuristically decided parameters, so that in the case of a large-scale problem, calculation time can be shortened by the use of smaller parameter values.

We can see from **Fig. 6** that MRS required more calculation time than any other method. MRS is a method for selecting (from training data) the data that meet specific conditions. For instance, MRS constructs 1-nearest neighboring method on such data and checks whether all training data can be classified. If all training data cannot be classified, the same checking procedures are repeated by increasing such data one by one until all training data can properly be classified. The repetitive checking procedures seem to have led to the longer calculation time required by MRS, relative to that required for other methods.

The shape of the line graph seems to suggest that the proposed method uses an exponential time algorithm. In the case of a large-scale problem, therefore, the search of all features could not be completed within a realistic time. Then, it would be necessary to first decrease the number of candidate features: some important features could first be selected by making them lower-dimensional through principal component analysis or auto encoder, or by applying existing methods that require a shorter calculation time (BW-Ratio, ReliefF, etc.). Therefore, the proposed method should be used only after the scale of a problem is reduced.

4. Conclusion

In recent years, many studies have been performed on the classification of human body motions using a combination of inertia sensors and machine learning. Such human body motion classifications require a huge amount of training data. Practically, however, it is often difficult to collect long-term data from a large number of subjects, both because subjects must be experienced at specific skills and because of concerns regarding fatigue and injuries. Therefore, many studies have constructed human body motion classifiers using training data collected from a small number of subjects [1–6]. In contrast, training data from a small number of subjects is susceptible to generalization errors caused by individual and trial differences. To suppress the generation of such generalization errors, we selected feature spaces from the training data that were less likely to generate generalization errors due to individual and trial errors. However, few studies were available on indices for evaluating such feature spaces.

This paper proposed evaluation indices for representing the likelihood that a feature space would generate generalization errors due to individual and trial differences. The proposed evaluation indices were obtained by first using acquired data to construct probability distributions representing individual and trial differences, and then by applying such probability distributions in calculating the risks of causing generalization errors. To verify the effectiveness of the proposed method, we acquired sensor data for butterfly and breaststroke swimming and selected feature spaces that were effective at classifying the swimming styles. In selecting such feature spaces, we applied the proposed method and certain existing methods (BW-Ratio [7], OOB [8], ReliefF [9], MRS [10]). We constructed a breaststroke/butterfly classifiers by applying SVM to the selected feature spaces. The accuracy verification conducted with test data verified that the feature space obtained using the proposed method had significantly higher mean F-measures than existing methods.

Furthermore, we measured and compared the calculation time for each method. We found that the proposed method required a longer calculation time than existing methods, and that the proposed method could have the characteristics of an exponential time algorithm. Although the proposed method can solve the problem assigned in Section 3 within a reasonable amount of time, in the case of a large-scale classification problem, it would need to use higher-dimensional feature spaces with an increased number of candidate features. Eventually, an enormous calculation time would be required. Practically, therefore, it would be useful to apply the proposed method only after reducing the problem scale by selecting effective features from all features for classification by existing methods that require short calculation times.

The use of the proposed evaluation indices enabled us to search for feature spaces that were less likely to cause generalization errors because of individual and trial differences, even in environments where a sufficient amount of training data could not be collected. This paper, however, only verified the effectiveness of the proposed method for one theme. In the future, therefore, we will need to further verify the effectiveness and generalization performance of the proposed method for more themes. With training data sizes as parameters, we will also need to verify the training data sizes from which the proposed method can search for effective features.

Acknowledgements

This work was supported in part by JSPS Grant-in-Aid for Young Scientists (B) (Grant No. 17K13179; Y. Omae and Grant No. 26800129; H. Takahashi) and by JSPS Grant-in-Aid for Scientific Research (C) (Grant Nos. 16K06156, 16K04672 and 17K05437; H. Takahashi).

References:

[1] A. M. Khan, Y. K. Lee, S. Y. Lee, and T. S. Kim, "A Triaxial Accelerometer-Based Physical-Activity Recognition via Augmented-Signal Features and a Hierarchical Recognizer," IEEE Tran. on Information Technology in Biomedicine, Vol.14, No.5, pp. 1166-1172, 2010.

[2] J. Lester, T. Choudhury, and G. Borriello, "A Practical Approach to Recognizing Physical Activities," Int. Conf. on Pervasive Computing, pp. 1-16, 2006.

[3] Z. He and L. Jin, "Activity Recognition from Acceleration Data Based on Discrete Consine Transform and SVM," IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 5041-5044, 2009.

[4] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner, "Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.28, No.10, pp. 1553-1567, 2006.

[5] P. Siirtola, P. Laurinen, J. Roning, and H. Kinnunen, "Efficient Accelerometer-Based Swimming Exercise Tracking," Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symp., pp. 156-161, 2011.

[6] Y. Kon, Y. Omae, K. Sakai, H. Takahashi, T. Akiduki, C. Miyaji, Y. Sakurai, N. Ezaki, and K. Nakai, "Toward Classification of Swimming Style by Using Underwater Wireless Accelerometer Data," ACM Int. Symp. on Wearable Computers, pp. 85-88, 2015.

[7] T. Hastie, R. Tibshirani, J. Friedman, M. Sugiyama, T. Ide, T. Kamishima, T. Kurita, and E. Maeda, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Kyoritsu Publisher, 2014.

[8] L. Breiman, "Random Forests," Machine Learning, Vol.45, No.1, pp. 5-32, 2001.

[9] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," Machine Learning, Vol.53, No.1-2, pp. 23-69, 2003.

[10] X. W. Chen and C. J. Jong, "Minimum Reference Set Based Feature Selection for Small Sample Classifications," The 24th Int. Conf. on Machine Learning, pp. 153-160, 2007.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. of the American Statistical Association, Vol.97, No.457, pp. 77-87, 2002.

[12] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," IEEE Trans. on Computational Biology and Bioinformatics, Vol.1, No.4, pp. 181-190, 2004.

[13] C. A. Ronao and S. B. Cho, "Human Activity Recognition Using Smartphone Sensors with Two-Stage Continuous Hidden Markov Models," 10th Int. Conf. on Natural Computation, pp. 681-686, 2014.

[14] Y. Omae, Y. Kon, M. Kobayashi, K. Sakai, A. Shionoya, H. Takahashi, T. Akiduki, K. Nakai, N. Ezaki, Y. Sakurai, and C. Miyaji, "Swimming Style Classification Based on Ensemble Learning and Adaptive Feature Value by Using Inertial Measurement Unit," J. of Advanced Computational Intelligence and Intelligent Informatics, Vol.21, No.4, pp. 616-631, 2017.

[15] A. Wang, G. Chen, J. Yang, S. Zhao, and C. Y. Chang, "A Comparative Study on Human Activity Recognition Using Inertial Sensors in a Smartphone," IEEE Sensors J., Vol.16, No.11, pp. 4566-4578, 2016.

[16] R. Akhavian, L. Brito, and A. Behzadan, "Integrated Mobile Sensor Based Activity Recognition of Construction Equipment and Human Crews," Conf. on Autonomous and Robotic Construction of Infrastructure, 2015.

[17] Y. Kon, Y. Omae, K. Sakai, H. Takahashi, T. Akiduki, C. Miyazi, Y. Sakurai, N. Ezaki, and K. Nakai, "Swimming Style Classification for Developing System of Swimming Performance and Technique Evaluation," JSME Symp.: Sports engineering and Human Dynamics 2015, A-15, 2015.

[18] Y. Ohgi, K. Kaneda, and A. Takakura, "A swimming style prediction using the chest acceleration," Symp. on Sports and Human Dynamics 2012, pp. 98-103, 2012.

[19] W. Choi, J. Oh, T. Park, S. Kang, M. Moon, U. Lee, I. Hwang, and J. Song, "Mobydick: An Interactive Multi-Swimmer Exergame," The 12th ACM Conf. on Embedded Network Sensor Systems, pp. 76-90, 2014.

[20] U. Jensen, F. Prade, and B. M. Eskofier, "Classification of Kinematic Swimming Data with Emphasis on Resource Consumption," Body Sensor Networks (BSN), 2013 IEEE Int. Conf., pp. 1-5, 2013.

[21] L. Bao and S. S. Intille, "Activity Recognition from User-Annotated Acceleration Data," Pervasive Computing, pp. 1-17, 2004.

[22] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity Recognition from Accelerometer Data," AAAI, Vol.5, pp. 1541-1546, 2005.

[23] M. Kobayashi, Y. Omae, Y. Kon, K. Sakai, A. Shionoya, H. Takahashi, Y. Sakurai, C. Miyaji, K. Nakai, K. Nakai, N. Ezaki, and T. Akiduki, "Analysis of Stroke Duration for Swimming Motion Coaching System by Using a Sensor Device," The Robotics and Mechatronics Conf. 2016 (ROBOMECH2016), 2A1-11b5-1-4, 2016.

[24] Sports Sensing Co., Ltd., "9 Axis Waterproof Type Wireless Motion Sensor," http://www.sports-sensing.com/products/motion/inertia/motionwp01.html, [accessed Feb. 9, 2016].

[25] Sony, Digital Video Camera HDR-CX720V, http://www.sony.jp/handycam/products/HDR-CX720V/spec.html, [accessed Feb. 17, 2016].

Name:
Yuto Omae

Affiliation:
Department of Electrical Engineering, National Institute of Technology, Tokyo College

Address:
1220-2, Kunugida, Hachioji, Tokyo 193-0942, Japan

Brief Biographical History:
2016.3 Dr. Eng. degree from Nagaoka University of Technology
2016.4-2017.3 Researcher, Japan Institute of Sports Sciences
2017.4- Assistant Professor, National Institute of Technology, Tokyo College

Main Works:
- "Method to Detect Change of Motivation to Enroll in University by Survey of Career Perceptions," J. of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol.27, No.5, pp. 743-756, 2015.

Membership in Academic Societies:
- Japan Society for Fuzzy Theory and Intelligent Informatics (J-SOFT)
- The Institute of Electronics, Information and Communication Engineers (IEICE)
- Japan Society for Educational Technology (JSET)

Name:
Hirotaka Takahashi

Affiliation:
Department of Information and Management Systems Engineering, Nagaoka University of Technology

Address:
1603-1, Kamitomioka, Nagaoka, Niigata 940-2188, Japan

Brief Biographical History:
2005.3 D.Sci. degree from Niigata University
2010.4-2011.3 Assistant Professor, Yamanashi Eiwa College
2011.4-2012.3 Lecturer, Yamanashi Eiwa College
2012.4-2013.3 Associate Professor, Yamanashi Eiwa College
2013.4- Associate Professor, Nagaoka University of Technology

Main Works:
- "Analysis of gravitational waves from binary neutron star merger by Hilbert-Huang transform," Physical Review D, Vol.93, 123010-1-11, 2016.

Membership in Academic Societies:
- Japan Society for Fuzzy Theory and Intelligent Informatics (J-SOFT)
- Operations Research Society of Japan (ORSJ)
- Physical Society of Japan (JPS)