Doctoral Thesis

Unlearned class estimation based on complementary event model and its application to biological signal classification

(余事象モデルに基づく 未学習クラス推定法の開発 と生体信号分類への応用)

Takayuki MUKAEDA

Department of Mathematics, Physics, Electrical Engineering and
Computer Science,
Graduate School of Engineering Science,
Yokohama National University

Advisory Professor Keisuke SHIMA

March 2021

Abstract

Computational automation systems based on machine learning have seen significant development in recent years, with examples including image recognition technology, man-machine interfaces and temporal data models for stock price prediction. In the fields of welfare and medicine, effectiveness in pattern recognition technology has also been reported for anomaly detection in medical imagery and the development of myoelectric prosthetic hands.

Classifiers in general simply identify classes predefined in training, but cannot be applied to consideration of specific undefined classes with abnormal patterns. If unknown patterns belonging to undefined classes are input into a trained classifier, misclassification relating to predefined classes will inevitably result. This can cause misclassification of unknown diseases in diagnosis assistance and critical errors in interface control. To address this, unlearned patterns not included in training data need to be recognized.

A variety of classifiers are used to detect unexpected outliers during training. The authors also performed research to enable highly accurate anomaly detection with the novel stochastic OVRGMN (One-vs.-Rest Gaussian Mixture Network) approach, with definition and application of complementary Gaussian distribution. The main issues of previous methods involve the difficulty of stable optimization with small pools of learning data and appropriate thresholds, along with the challenges involved in setting appropriate empirical thresholds. The OVRGMN is also premised on application of the static characteristics of input data, and cannot be used to handle time-series data.

This research was performed to develop novel unlearned class detection superior to previous methods and support classification and evaluation of biological signals.

The Normal and Complementary Gaussian Mixture Network (NACGMN) was applied as a novel probabilistic neural approach with unlearned class detection for high classification performance and stable training even with small training

samples. The NACGMN incorporates Gaussian mixture models (GMMs) and complementary Gaussian mixture models (CGMMs) representing distribution of training and unlearned classes, respectively. Since the parameters of both distributions can be determined as weighting coefficients of the network with relaxed statistical constraints, the NACGMN supports more stable training. The outcomes of classification experiments employing artificial data and EMG signals demonstrated the validity of the NACGMN.

To extend the proposed network to a classifier capable of handling time-series information, novel One-vs.-Rest hidden Markov Model (OVRHMM) classification was proposed in which time-series data for trained classes are modeled using hidden Markov models (HMMs) with GMMs and unlearned patterns are expressed using HMMs with CGMMs. Based on *a posterior* probabilities estimated using Bayes' theorem, the OVRHMM can be used to stochastically evaluate the degree of abnormality for input signals. Experimental results indicated that the ORVHMM enabled classification of time-series data that cannot be discriminated using static classifiers.

A time-series data classification model based on the Hidden Semi-Markov Model (HSMM), in which an anomaly state with the CGMM is introduced, was also proposed for more detailed anomaly detection. The related novel classification method combining Bayesian discrimination in consideration of temporal information and estimation of state transition sequencing supports highly accurate classification and detailed anomaly detection. To evaluate the effectiveness of the proposed HSMM for real data, the approach was applied to the classification of care-worker motion with outcomes demonstrating accurate detection and recognition of important work among innumerable movements.

The above achievements show that unlearned-class detection based on complementary event models enabled stable training independent of learning conditions and high-precision classification in consideration of time-series data.

あらまし

　近年，機械的な自動化を目的として機械学習を利用した画像認識システムやマンマシンインタフェース，数値予測が可能な時系列データモデルなどの開発が盛んに行われている．医療・福祉分野では医療画像に対する異常検知や生体信号による制御が可能な電動義手の開発など様々な課題に対してパターン識別法が適用され，その有効性が確認されている．

　一般的なパターン識別問題では識別モデルの生成のために教師データが存在し，予め識別対象クラスのサンプルデータやクラスの数を与えることで識別器の学習を実現する．しかしながら，事前に全ての起こりうるパターンを網羅する学習を行うことは極めて困難であり，学習対象としていないパターンが入力された場合には定義されていない事象を適切に分類できず，必然的に誤識別が誘発されてしまう．この問題は病症の診断支援における未知の症例の誤識別やインタフェースの制御における想定しない操作ミスなどを誘発する可能性があり，事前の学習に含まれていない未学習のパターンを推定する機構が必要不可欠である．

　未学習のパターンは学習時に想定しない異常値として捉えることができ，扱うデータの性質や学習時の条件などに応じた様々な手法が提案されている．我々の研究グループにおいても未学習クラスの分布を仮想的に表現する余事象分布を導入した新たな確率モデル (OVRGMN: One-Vs-Rest Gaussian Mixture Network) によって高精度な異常検知を可能にした．一方，これらの従来法の主要な課題として，学習データが少ない場合に安定的な学習が困難になることや異常の有無を判断する適切な閾値設定が難しいことなどが挙げられる．また，OVRGMN においては入力データの静的特徴の利用を前提としており，時間とともに変動する時系列データを取り扱うことはできない．データの時間情報を利用可能な識別手法が高い分類性能を有することが広く知られており，時系列データへの適用が求められる．

　本論文では従来の課題を解消する新たな未学習クラス推定法および学習則を提案し，生体信号の分類・評価へ応用することである．

　まず少ない学習データからも高精度な分類および安定的な学習を実現するために，新たな未学習クラス推定確率ニューラルネット (Normal And Complementary

Gaussian Mixture Model: NACGMN) を提案した．NACGMN は学習クラスの分布を表現する混合正規分布 (Gaussian Mixture Model: GMM) と未学習クラスの分布を仮定した混合余事象分布 (Complementary GMM: CGMM) を内包し，両分布のパラメータを統計的制約が緩和されたネットワークの重み係数として獲得できるため，より安定的な学習を実現できる．実験では人工データや筋電位信号の分類を行い，学習データ数が少ない場合においても従来法と比較して優れた識別精度が得られることが示された

　　次に，提案法を時系列情報を考慮可能な分類モデルへ拡張するための基礎的検討を行い，データの時間変化を状態遷移により表現する隠れマルコフモデル (Hidden Markov Model: HMM) に基づく新たな分類モデル (One-Vs-Rest HMM: OVRHMM) を提案した．OVRHMM では事前に想定しない時系列データを CGMM を有する HMM によりモデル化することで，入力信号に対する異常度を確率的に評価できる．評価実験の結果から従来の静的な識別器では分類不可能な時系列データを ORVHMM により分類できることが示された．また，足圧データによる神経変性疾患の検知問題において，OVRHMM は健常者データのみを学習し，異常を含む患者のデータを未学習クラスに識別することで，両データを学習に用いた従来法と同等の精度で歩行時の異常を検出できた．

　　さらに，適用範囲の拡大を目的として OVRHMM の機能を拡張し，時系列データの含まれる瞬間的な異常も検知可能な時系列データ分類モデルを検討した．提案法では隠れセミマルコフモデル (Hidden Semi-Markov Model: HSMM) の内部に異常値の分布を表現する未分類状態を導入し，時間情報を考慮したベイズ識別と状態遷移列推定を組み合わせた新たな識別法により高精度な多クラス分類と詳細な異常検知の両方を実現する．評価実験では実際の介護施設で計測した介護士の姿勢情報に対して提案法を適用し，無数に存在する動作の中で重要な要素作業のみを検出し，分類できるか検討した．結果では慣性センサから推定した簡易な姿勢情報からも高精度に特定作業を検知できる可能性が示唆され，提案法の有効性を確認した．

　　以上のことから，提案した余事象モデルに基づく未学習クラス推定法によって従来法の課題であった学習条件に依存しない安定的な学習や時系列データの時間変化を考慮した高精度な分類を実現できたことが示された．

Acknowledgements

First of all, I would like to show my greatest appreciation to my supervisor, associate professor Keisuke Shima for leading me to this very exciting area. He supported and encouraged me for about 5 years, I was able to learn a lot of thing in his laboratory. The experiences I have learned in his laboratory have changed my life. I had a happy opportunity to meet pattern classification and biological signal processing. I am filled with gratitude.

I also appreciate the feedback offered by professor Kohno, professor Hamagami, professor Ochiai, associate professor and associate professor Nakata. To all the members of Shima Laboratory, I would like to express my appreciation.

Also, I am deeply grateful to associate professor Takayuki Tanaka and associate professor Hiroyuki Izumi for helpful discussions and comments for my research. Without their supports, this dissertation could not been completed.

Finally, I would like to express my sincere gratitude to my family for giving me warm encouragements and the opportunity to pursue my doctor's degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Purpose

Computational automation systems based on machine learning have seen significant development in recent years, with examples including image recognition technology, man-machine interfaces and temporal data models for stock price prediction [1]. In the felds of welfare and medicine, the effectiveness of pattern recognition technology has also been reported for anomaly detection in medical imagery [2] and the development of myoelectric prosthetic hands [3–6]. A number of advanced approaches have also been proposed for the accurate pattern recognition algorithms required by these systems toward biological signal classifcation [4, 7]. Classifers in general simply identify classes predefined in training, but cannot be applied to consideration of specifc undefined classes with abnormal patterns. If unknown patterns belonging to undefined classes are input into a trained classifer, misclassification relating to predefined classes will inevitably result. This can cause misclassification of unknown diseases in diagnosis assistance and critical errors in interface control.

To address this issue, unlearned patterns not included in training data need to be recognized. A variety of classifiers are used to detect unexpected outliers during training, and related methods can be classified roughly in terms of data characteristics and training conditions. In this regard, pattern recognition tools

based on temporal data, such as the hidden Markov model (HMM) [7], have been applied to handle time-series data. Image classification tools such as convolutional neural networks [8, 9] and variational autoencoders [15] have also demonstrated effectiveness in abnormal-point detection.

Although extensive normal data are observed in anomaly detection problems, the number of abnormal values that can be applied in learning is generally very limited. As standard learning regulations involve consideration of discrimination accuracy for overall imbalanced data as described above, discrimination accuracy for rare outliers may be neglected [12]. To address this, oversampling for outliers [11, 12], undersampling for the majority [12] and suppression of misidentification for minorities based on boosting have been proposed [13]. Lu *et al.* developed electrocardiogram (ECG) classification with random oversampling to address imbalanced data problems, with results from application to the MIT-BIH Arrhythmia Database and MIT-BIH ST Change Database demonstrating the approach as a promising alternative superior to state-of-the-art methods. In addition, while many normal data are observed in anomaly detection problems, the number of abnormal values that can be applied in learning is generally very limited.

To tackle anomaly detection issues in which abnormal values are occasionally observed, semi-supervised learning considerations such as Mahalanobis distance [16], single-class support vector machine (SVM) [17,18], isolation forest [19] and single-class neural network [20] approaches are often used, as these do not require training data for anomalies. However, these methods involve difficulty in determining appropriate thresholds and do not guarantee both high-accuracy multiclass discrimination and anomaly detection. To address this, the authors previously proposed a probabilistic classifier based on a novel Gaussian mixture model (GMM) that can be used to estimate *a posteriori* probability for unlearned classes [21]. Although this approach can be used for a variety of problems, model parameter optimization is challenging with training data sets. The static nature of these approaches also makes them incompatible with time-series data analysis.

Against such a background, this research was performed to support the development of novel unlearned class detection superior to previous methods and enable the application of classification and evaluation for biological signals. The content of this paper is summarized below.

- **Improved parameter optimization for small datasets**

    Unlearned class determination based on the novel stochastic One-vs.-Rest Gaussian Mixture Network (OVRGMN) proposed by Shima [21] supports probabilistic evaluation of anomalies in input data and enables multi-class classification and anomaly detection with single classifiers. Although the effectiveness of this approach has been confirmed in previous studies, stable learning is difficult when training datasets are small as with other conventional methods. The main causes of learning instability are the inability to calculate the inverse of a covariance matrix due to statistical constraints and the difficulty of learning for multiple components. Accordingly, this paper proposes novel complementary Gaussian distribution (CGD) and parameter transformation based on Log-linearization techniques [22] for relaxation of statistical constraints in parameter estimation. Thus, learning stability is improved for small data sets by expanding probability calculation using deformed parameters for a neural network structure.

- **Probabilistic evaluation of abnormalities for time-series data**

    To expand the rather limited range of application for the proposed method (in which only the static characteristics of data are considered) to the classification of actual data, unlearned class estimation based on a hidden Markov model (HMM) is proposed for time-series data. The HMM enables probabilistic anomaly determination independent of threshold settings as well as semi-supervised learning without the need for outlier learning data. The approach was applied to biological signal classification and evaluated for applicability to medical diagnosis.

- **Detailed anomaly detection with unlearned sequential pattern recognition**

Abnormalities in time-series data include unlearned patterns as well as spike noise and sudden trend changes. The proposed HMM can be used to evaluate the degree of abnormality for an entire input signal, but may not fully detect abnormal points or exhibit sufficient discrimination performance depending on the magnitude and extent of the abnormalities involved. Hence, this paper proposes a novel pattern recognition model based on the hidden semi-Markov model (HSMM) [23, 24] with an anomaly state for detailed anomaly detection. The classification combines Bayesian discrimination in consideration of temporal information and estimation of state transition sequencing, thereby enabling highly accurate classification and detailed anomaly detection.

1.2 Related Work

1.2.1 Semi-supervised anomaly detection with static features

The many anomaly detection methods previously proposed include that of Domingues *et al.*, which involves categorization of probabilistic, distance-based, information theory, neural network-based, domain-based and isolation-based methods [25]. The probabilistic method includes the robust kernel density estimator (RKDE) proposed by Kim *et al.* [26] and the probabilistic principal component analysis (PPCA) proposed by Tipping *et al.* [27]. For the isolation-based method, Liu *et al.* proposed an isolation forest (IF) tree structure-based classifier and demonstrated effectiveness in detecting anomalies with no teacher label [19]. However, these approaches to semi-supervised learning premise two-class discrimination to evaluate input data normality, therefore requiring multiple classifiers to achieve both multi-class identification and anomaly detection. In addition, evaluating only the degree of abnormality often requires the setting of threshold values, with complex empirical results and training for multiple classifiers poten-

tially hindering classification accuracy. The OVRGMN approach proposed by Shima *et al.* helps to address these issues, but separate regulations are required for the learned and unlearned classes.

This paper proposes a novel classification approach enabling multi-class discrimination and anomaly detection with a single classifier.

1.2.2 Conventional GMM parameter determination

The statistical parameters of probability density functions (PDFs) must be estimated to determine target-class training sample distribution. For the GMM, typical estimation involves the k-means algorithm [21], the expectation-maximization (EM) algorithm [28], variational training [29], Markov chain Monte Carlo (MCMC) methods [30] and infinite GMM application [31]. The k-means algorithm and EM algorithm are popular in parameter estimation because they are simpler and faster than other algorithms. It is also widely known that the variational training and MCMC methods enable highly accurate training data distribution, but require a large body of learning data for high-precision estimation and are sub-optimal for determination of the covariance matrix for training data with small variances. In addition, training in the MCMC and infinite GMM approaches also involves a significant time burden.

Meanwhile, if the probabilistic models determined are used for multiclass classification, accuracy is more important than in estimation for distribution parameters. To address this issue, Tsuji *et al.* proposed the Log-Linearized Gaussian Mixture Network (LLGMN) approach [22] involving parameter transformation to convert GMM statistical parameters (such as mean vector and covariance matrix values) into weighting coefficients without statistical constraints. This enables handling of the GMM as a stochastic neural network, and parameter estimation in consideration of classification accuracy can be realized using error backpropagation. As the number of estimated parameters can be reduced in this approach, highly accurate classification can be achieved even with a small body of data. However, this technique cannot be applied with the unlearned class estimation

model reported by Shima *et al.* [21]. In the LLGMN, statistical parameters also cannot be extracted from the acquired weighting factors, and readability in its role as a generative model is impaired.

This paper proposes novel training based on log-linearization in consideration of unlearned class data and stochastic parameter extraction from the weight coefficients determined.

1.2.3   Anomaly detection for time-series data

Abnormalities for time-series data include unlearned patterns as well as spike noise and sudden changes in trends, with methods based on data characteristics and training conditions [32, 33]. Most previously proposed anomaly detection methods have featured a classifier specialized for the data [34]. Bae *et al.* reported on classification allowing the identification of unusual gait phase transition and absent gait phases in subjects with Parkinson's disease based on ground reaction [36]. Fuse *et al.* proposed abnormal-period detection based on a sticky hierarchical Dirichlet-process hidden Markov modeling (sHDP-HMM) [35]. In this method, two sHDP-HMMs were trained using normal data and test data involving anomalies, respectively, and comparing the structure of the two outcomes enabled anomaly detection and modeling. The proposed method was applied to mesh population data, and demonstrated accurate detection performance for anomalous temporal population changes.

The unlearned detection proposed by Shima *et al.* does not utilize temporal information relating to time-series data, but features extensibility to time-series data models. The introduction of probabilistic models of this kind (such as the HMM and HSMM) may enable unlearned class detection for time-series data. The HSMM can be used for detailed modeling of changes in time-series data and has a wide range of applications, making it ideal for the development of a highly versatile anomaly detection approach.

Against such a background, this paper proposes novel unlearned pattern recognition utilizing temporal information on input signals and a related training al-

gorithm.

1.3   Dissertation Outline

This paper proposes novel unlearned class detection based on a complementary event model for stable unsupervised training with small datasets and pattern recognition utilizing temporal information from time-series data. To demonstrate related performance, the approach was applied to biological signal classification. The structure of the paper is outlined below.

Section 2 details the novel probabilistic neural Normal and Complementary Gaussian Mixture Network (NACGMN), which can be used to estimate *a posterior* probability for unlearned classes. The network can be optimized using a back-propagation algorithm and handle semi-supervised learning in which only normal training data with no abnormal samples are used. The method does not require trial-and-error determination for thresholds, and enables multi-class classification and anomaly detection with a single network. Parameter extraction for the NACGMN was also developed with focus on readability improvement, and two pattern recognition experiments were performed to evaluate the approach. For artificial data, the NACGMN was found to support superior classification with a small training sample set. The validity of the proposed parameter extraction was also confirmed from the decision region of estimated probability models. In motion classification experiments, the approach was applied to recognition of EMG signals from seven subjects each performing eight forearm motions. The results showed that the method produced significantly higher classification accuracy for all subjects.

Section 3 outlines the novel One-vs.-Rest Hidden Markov Model (OVRHMM) as a prototype for sequential pattern recognition to extend the proposed network to a classifier capable of handling time-series information. The model allows identification of unexpected classes in the learning process based on time-series data characteristics. In the OVRHMM, time-series data for trained classes are pro-

duced using hidden Markov models (HMMs) with GMMs, and unlearned patterns are expressed using HMMs with CGMMs. Based on *a posterior* probabilities estimated using Bayes' theorem, the OVRHMM can be used to stochastically evaluate degrees of abnormality for input signals. Experiment results indicated that the model enabled classification of time-series data that static classifiers were unable to discriminate. Here, three pattern recognition experiments employing artificial signals, EMG signals and foot pressure data from ambulation were performed to validate the performance of the approach. In neurodegeneration evaluation experiments based on gait data, subjects were placed in the unlearned class for training without disorder gait data. The OVRHMM was found to demonstrate a level of performance similar to that of the previous method, in which disorder data are used for training.

Section 4 details development of the OVRHMM and outlines novel HSMM-based sequential pattern recognition incorporating unlearned states to represent unexpected anomalies in the learning process based on time-series data characteristics. The newly developed classification method, combining Bayesian discrimination in consideration of temporal information and state transition sequence determination, provides high classification performance with focus on time-series information relating to signals, and allows more detailed anomaly detection than the OVRHMM. Application for classification of care-worker motion to evaluate the effectiveness of the proposed HSMM for real data demonstrated accurate detection and recognition of important actions among innumerable movements.

Section 5 concludes the paper, outlining a number of remaining challenges and plans for future work.

Chapter 2

A novel probabilistic neural network with unlearned-class detection based on normal and complementary Gaussian mixture models

2.1 Introduction

Chapter 2 involves the proposal of a novel probabilistic neural network (referred to as a normal and complementary Gaussian mixture network, or NACGMN) that can be used to handle multiclass classification and anomaly detection. The network incorporates a GMM and a newly defined complementary GMM (CGMM), and allows optimization of statistical model parameters as network coefficients. With a unique error backpropagation algorithm, the method allows accurate multiclass classification and anomaly detection.

In this chapter, Section 2.2 here outlines the incorporated probability function for unlearned classes and multi-class classification based on Bayes' theorem

in consideration of the unlearned classes involved, Section 2.3 describes the structure of the proposed neural network and related learning methods, Section 2.4 covers performance evaluation using artificial data and discusses application to electromyogram (EMG) discrimination and NACGMN validity, and Section 2.5 draws conclusions and outlines future study plans.

2.2 GMM- and CGMM-based pattern recognition

2.2.1 Bayesian estimation based on Gussian mixture models

The GMM represents probability distribution as the linear sum of a finite number of Gaussian distributions, and demonstrates high expressiveness even for data with complex distributions. A GMM with K classes is used to express the probability density functions (PDF) of a D-dimensional feature vector $\boldsymbol{x}_n \in \Re^D (n = 1, \cdots, N)$ as

$$f(\boldsymbol{x}_n) = \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} \alpha_{k',m'} g\left(\boldsymbol{x}_n; k', m'\right) \tag{2-1}$$

$$g\left(\boldsymbol{x}_n; k, m\right) = (2\pi)^{-\frac{D}{2}} \left|\boldsymbol{\Sigma}_{k,m}\right|^{-\frac{1}{2}} \exp\left\{q(\boldsymbol{x}_n; k, m)\right\} \tag{2-2}$$

$$q(\boldsymbol{x}_n; k, m) = -\frac{1}{2} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{k,m}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{k,m}^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_{k,m}\right), \tag{2-3}$$

where $g\left(\boldsymbol{x}_n; k, m\right)$ represents normal distribution, N is the number of training samples, M_k is the number of Gaussian components in each event, and each component has the mixture coefficient $\alpha_{k,m} > 0$, the mean vector $\boldsymbol{\mu}_{k,m} \in \Re^D$ and the covariance matrix $\boldsymbol{\Sigma}_{k,m} \in \Re^{D \times D}$. Here, $\sum_{k=1}^{K} \sum_{m=1}^{M_k} \alpha_{k,m} = 1$ is met. Based on Bayes' theorem, the a posteriori probability of individual classes $p(k|\boldsymbol{x}_n)$ ($k = 1, \cdots, K$) is implemented with

$$p(k|\boldsymbol{x}_n) = \frac{1}{f(\boldsymbol{x}_n)} \sum_{m=1}^{M_k} \alpha_{k,m} g\left(\boldsymbol{x}_n; k, m\right). \tag{2-4}$$

Based on the assumption that multiple events do not occur simultaneously, selection of the class with the maximum $p(k|\boldsymbol{x}_n)$ value reaches the classification of $\boldsymbol{x}_n$ in a K class.

2.2.2 Unlearned class detection using complementary GMMs

This section discusses probabilistic anomaly detection premised on novel PDF-based complementary Gaussian mixture models (CGMMs). In Bayes' theorem (Eq. 2–4), the *a posteriori* probability of each class is estimated from the likelihood ratio of predefined K classes. Hence, general GMMs cannot be used to identify unlearned classes that do not belong to a particular class. Accordingly, a $K+1$ classification problem based on the unlearned class ($k = 0$) was considered. Shima *et al.* assumed that anomaly data do not belong to a Gaussian component $\{k, m\}$ of training samples, and estimated the occurrence probability of data belonging to other components based on the following PDF $h_{\mathrm{pr}}(\boldsymbol{x}_n; \boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m}, \xi_{k,m})$ [21]:

$$h_{\mathrm{pr}}\left(\boldsymbol{x}_n; k, m\right) = (2\pi)^{-\frac{D}{2}} \left|\boldsymbol{\Sigma}_{k,m}\right|^{-\frac{1}{2}} \left(\xi_{k,m}^{\frac{D}{2}} - 1\right)^{-1}$$
$$\times \left[\exp\left\{\xi_{k,m}^{-1} q(\boldsymbol{x}_n; k, m)\right\} - \exp\left\{q(\boldsymbol{x}_n; k, m)\right\}\right] \qquad (2\text{–}5)$$

Here, $\xi_{k,m} > 1$ is a parameter contributing to the variance of h_{pr}, and $\exp\{\xi_{k,m}^{-1} q(\boldsymbol{x}_n)\} \geq \exp\{q(\boldsymbol{x}_n)\}$ must be met. However, h_{pr} parameter estimation for multiple components with the PDF is extremely challenging due to the above statistical constraints.

Based on the above, the authors propose complementary Gaussian distribution (CGD) expressed by the product of a multivariate quadratic function and normal distribution. The proposed PDF $h(\boldsymbol{x}_n; \boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m}, \varepsilon_{k,m})$ is defined as

$$h\left(\boldsymbol{x}_n; k, m\right) = -2D^{-1}(2\pi)^{-\frac{D}{2}} \varepsilon_{k,m}^{-1} \left|\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}\right|^{-\frac{1}{2}}$$
$$\times q(\boldsymbol{x}_n; k, m) \exp\left\{\varepsilon_{k,m}^{-1} q(\boldsymbol{x}_n; k, m)\right\}. \qquad (2\text{–}6)$$

Here, h is defined as a PDF satisfying $\int_{-\infty}^{\infty} h(\boldsymbol{x})d\boldsymbol{x} = 1$, and $\varepsilon_{k,m} > 0$ is met. Here, the likelihood of an unlearned class can be estimated from a CGMM expressed

using the linear sum of h. GMM and CGMM examples are shown in Figure 2.1. Based on the above assumptions, occurrence probability $F(\boldsymbol{x}_n)$ relating to the unlearned class is calculated from a mixture model involving the GMM and CGMM, and is represented as

$$F(\boldsymbol{x}_n) = p(k \neq 0) \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} \alpha_{k',m'} g\left(\boldsymbol{x}_n; k', m'\right)$$

$$+ p(k = 0) \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} \beta_{k',m'} h\left(\boldsymbol{x}_n; k', m'\right) \qquad (2\text{--}7)$$

$$p(k \neq 0) = 1 - p(k = 0) \qquad (2\text{--}8)$$

$$\sum_{k=1}^{K} \sum_{m=1}^{M_k} \alpha_{k,m} + \sum_{k=1}^{K} \sum_{m=1}^{M_k} \beta_{k,m} = 1 \qquad (2\text{--}9)$$

Here, each complementary Gaussian component has the mixture coefficient $\beta_{k,m}$, and $p(k = 0)$ represents the *a priori* probability of the unlearned class. In $K + 1$ classification based on Bayes' theorem, the *a posteriori* probability for the feature vector $\boldsymbol{x}_n$ is estimated as

$$p(k|\boldsymbol{x}_n) = \begin{cases} \dfrac{p(k \neq 0)}{F(\boldsymbol{x}_n)} \displaystyle\sum_{m=1}^{M_k} \alpha_{k,m} g(\boldsymbol{x}_n; k, m) & (k \neq 0) \\ \dfrac{p(k = 0)}{F(\boldsymbol{x}_n)} \displaystyle\sum_{l=1}^{K} \sum_{m=1}^{M_l} \beta_{l,m} h(\boldsymbol{x}_n; l, m) & (k = 0). \end{cases} \qquad (2\text{--}10)$$

Based on Eq. (2–10), pattern classification with unlearned-class consideration can be achieved if the parameter set

$$\boldsymbol{\theta} = \left\{ p(k = 0), \alpha_{k,m}, \beta_{k,m}, \boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m}, \varepsilon_{k,m} \right\}_{\substack{k=1,\cdots,K, \\ m=1,\cdots,M_k}}$$

is adequately optimized with training samples.

2.2.3 Parameter conversion based on log linearization

For the GMM and CGMM, $\boldsymbol{\theta}$ values maximizing likelihoods for training samples must be estimated for each class. However, $\boldsymbol{\theta}$ involves statistical parameters
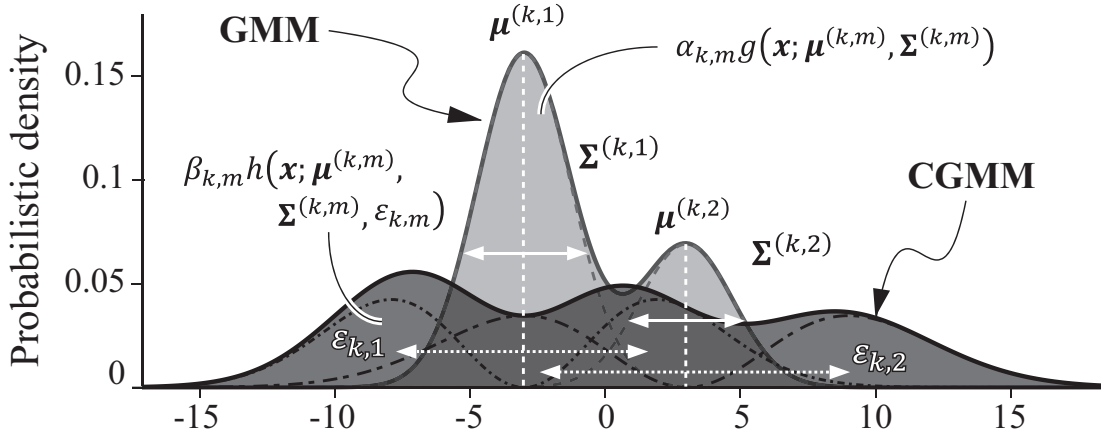
Fig. 2.1: GMM and CGMM examples

such as $\boldsymbol{\mu}_{k,m}$ and $\boldsymbol{\Sigma}_{k,m}$, as well as statistical constraints that make learning difficult. As $\boldsymbol{\Sigma}_{k,m}$ has to satisfy positive-semidefinite matrices, the determination of appropriate values with few training samples is problematic, and appropriate optimization of $\varepsilon_{k,m}$ requires a wealth of training data [21]. Although various training method of GMM has been developed so far, k-means algorithm and EM algorithm [28] has a problem that the calculation of *a posterior* probabilities becomes unstable, and the MCMC methods [30] requires a huge amount of learning time and training data. In addition, most of those training methods of GMM cannot apply to the optimization of CGMM. Accordingly, to deal with those problems, the proposed model parameter $\boldsymbol{\theta}$ is converted into weight coefficients with eased restrictions using the log-linearization approach developed by Tsuji *et al.* [5]. The log-linearization enables to avoid the inverse matrix operations of $\boldsymbol{\Sigma}_{k,m}$, and solve the problem of the EM algorithm.

First, the mean vector and the precision matrix for GMMs is defined as

$$\boldsymbol{\mu}_{k,m} = \left[\mu_1^{(k,m)}, \cdots, \mu_D^{(k,m)}\right]^{\mathrm{T}}, \ \ \boldsymbol{\Sigma}_{k,m}^{-1} = \left[s_{i,j}^{(k,m)}\right]_{\substack{1 \le i \le D \\ 1 \le j \le D}}.$$

and the multivariate quadratic function $q(\boldsymbol{x}_n; k, m)$ in Eq. (2–3) is given as

$$q(\boldsymbol{x}_n; k, m) = \boldsymbol{w}^{(k,m)} \boldsymbol{X}_n \qquad (2\text{–}11)$$

Here, $\boldsymbol{X}_n \in \Re^H$ and $\boldsymbol{w}^{(k,m)} \in \Re^H$ are defined as

$$\boldsymbol{X}_n = \left[1, \boldsymbol{x}^{\mathrm{T}}, x_1^2, x_1 x_2, \cdots, x_1 x_D, x_2^2, x_2 x_3, \cdots, \right.$$

$$\left. x_2 x_D, \cdots, x_D^2 \right]^{\mathrm{T}} \tag{2-12}$$

$$\boldsymbol{w}^{(k,m)} = \left[w_0^{(k,m)}, w_1^{(k,m)}, \cdots, w_{H-1}^{(k,m)}\right] \tag{2-13}$$

$$= \left[w_0^{(k,m)}, \sum_{j=1}^{D} s_{j,1}^{(k,m)} \mu_j^{(k,m)}, \cdots, \sum_{j=1}^{D} s_{j,D}^{(k,m)} \mu_j^{(k,m)}, \right.$$

$$-\frac{1}{2} s_{1,1}^{(k,m)}, -s_{1,2}^{(k,m)}, \cdots, -s_{1,D}^{(k,m)}, \cdots,$$

$$\left. -\frac{1}{2}(2 - \delta_{j,l}) s_{j,l}^{(k,m)}, \cdots, -\frac{1}{2} s_{D,D}^{(k,m)} \right] \tag{2-14}$$

$$w_0^{(k,m)} = -\frac{1}{2} \sum_{j=1}^{D} \sum_{l=1}^{D} s_{j,l}^{(k,m)} \mu_j^{(k,m)} \mu_l^{(k,m)} \tag{2-15}$$

where $\delta_{i,j}$ represents the Kronecker delta ($\delta_{i,j} = 1$ when $i = j$ and $\delta_{i,j} = 0$ when $i \neq j$) and the dimension H – a new feature vector transformed non-linearly with $\boldsymbol{X}_n$ – is $H = 1 + D(D+3)/2$. With Eq. (2–11), the *a posteriori* probability for each component of the GMM and CGMM, $Y_n^{(k,m)}$ and $Z_n^{(k,m)}$, can be given as

$$Y_n^{(k,m)} = \log\left\{\alpha_{k,m} g\left(\boldsymbol{x}_n; k, m\right)\right\}$$

$$= \boldsymbol{w}^{(k,m)} \boldsymbol{X}_n + w_{\mathrm{N}}^{(k,m)} \tag{2-16}$$

$$Z_n^{(k,m)} = \log\left\{\beta_{k,m} h\left(\boldsymbol{x}_n; k, m\right)\right\}$$

$$= w_\varepsilon^{(k,m)} \boldsymbol{w}^{(k,m)} \boldsymbol{X}_n + \log\left(-\boldsymbol{w}^{(k,m)} \boldsymbol{X}_n\right) + w_{\mathrm{C}}^{(k,m)}. \tag{2-17}$$

Here, $w_{\mathrm{N}}^{(k,m)}$, $w_{\mathrm{C}}^{(k,m)}$ and $w_\varepsilon^{(k,m)}$ are defined as

$$w_{\mathrm{N}}^{(k,m)} = \log \alpha_{k,m} - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_{k,m}| \tag{2-18}$$

$$w_{\mathrm{C}}^{(k,m)} = \log 2 + \log \beta_{k,m} - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_{k,m}|$$

$$- \left(1 + \frac{D}{2}\right) \log \varepsilon_{k,m} \tag{2-19}$$

$$w_\varepsilon^{(k,m)} = \varepsilon_{k,m}^{-1}. \tag{2-20}$$

Equations (2–16) and (2–17) enable concise probabilistic calculation utilizing a matrix product of transformed input $\boldsymbol{X}_n$ and a coefficient vector $\boldsymbol{w}^{(k,m)}$. Based

on this definition, $\exp(Y_n^{(k,m)})$, $\exp(Z_n^{(k,m)})$ are equal to the occurrence probability calculated from each component of the GMM and CGMM, respectively. Hence, the *a posteriori* probability in consideration of unlearned classes can be redefined based on Eq. (2–10) as

$$p(k|\boldsymbol{x}_n) = \begin{cases} \dfrac{p(k\neq 0)}{F_{\mathrm{LL}}(\boldsymbol{x}_n)} \displaystyle\sum_{m=1}^{M_k} \exp\left(Y_n^{(k,m)}\right) & (k \neq 0) \\ \dfrac{p(k=0)}{F_{\mathrm{LL}}(\boldsymbol{x}_n)} \displaystyle\sum_{l=1}^{K}\sum_{m=1}^{M_l} \exp\left(Z_n^{(l,m)}\right) & (k = 0) \end{cases} \tag{2–21}$$

$$F_{\mathrm{LL}}(\boldsymbol{x}_n) = p\left(k\neq 0\right) \sum_{l=1}^{K}\sum_{m=1}^{M_l} \exp\left(Y_n^{(l,m)}\right)$$

$$+ p(k=0) \sum_{l=1}^{K}\sum_{m=1}^{M_l} \exp\left(Z_n^{(l,m)}\right) \tag{2–22}$$

With the above transformations, the structure of a parameter set $\boldsymbol{\theta}$ transmutes to

$$\boldsymbol{\theta}' = \left\{ p(k=0), \boldsymbol{w}^{(k,m)}, w_{\mathrm{N}}^{(k,m)}, w_{\mathrm{C}}^{(k,m)}, w_{\varepsilon}^{(k,m)} \right\}_{\substack{k=1,\cdots,K,\\ m=1,\cdots,M_k}}$$

Optimization of $\boldsymbol{\theta}'$ supports pattern classification based on GMM and CGMM operation. Here, $\boldsymbol{\theta}'$ allows relaxation of the statistical constraints of $\boldsymbol{\theta}$ and optimization based on sequential training as typified by error backpropagation algorithm usage. Accordingly, the log-linearization operations enable to optimize both GMM and CGMM in parallel based on a assumption that normal training data should not be misclassified into the unlearned class.

2.3　Network model incorporating the GMM and CGMM

2.3.1　Network structure

　Figure 2.2 shows the structure of the proposed neural network incorporating the GMM and CGMM. The feedforward network contains five layers and has

weight coefficients in the second layer. First, the input feature vector $\boldsymbol{x} \in \Re^D$ is preprocessed and converted into the modified input vector $\boldsymbol{X} \in \Re^H$ based on Eq. (2–12). The first layer consists of H units corresponding to the dimensions of $\boldsymbol{X}$. The input-output relationship of the first layer is defined as

$$^{(1)}I_h = X_h \tag{2–23}$$

$$^{(1)}O_h = {}^{(1)}I_h \tag{2–24}$$

Here, $^{(1)}I_h$ and $^{(1)}O_h$ denote the input and an output of the hth unit, respectively.

The second layer consists of $\sum_{k=1}^{K} M_k$ training target class units, and is represented by Eq. (2–11). Each unit receives output weighted by $\boldsymbol{w}^{(k,m)}$ as

$$^{(2)}I_{k,m} = \sum_{h=1}^{H} {}^{(1)}O_h w_h^{(k,m)} \tag{2–25}$$

$$^{(2)}O_{k,m} = {}^{(2)}I_{k,m}. \tag{2–26}$$

Here, $^{(2)}I_{k,m}$ and $^{(2)}O_{k,m}$ denote the input-output of the training target classes.

The third layer consists of target class $\{k,m\}$ and unlearned class $\{\overline{k,m}\}$ units, and involves the execution of Eqs. (2–16) and (2–17) with $^{(2)}O_{k,m}$, the bias terms $w_{\mathrm{N}}^{(k,m)}$ and $w_{\mathrm{C}}^{(k,m)}$, and the coefficient $w_{\varepsilon}^{(k,m)}$. To simplify the network structure, Eq. (2–17s) was divided into two units: one involving calculation of the first and third terms, and the other implementing calculation of the second term. Thus, the input-output relationship of the third layer is defined as

$$^{(3)}I_{k,m} = {}^{(2)}O_{k,m} + w_{\mathrm{N}}^{(k,m)}, \qquad\qquad {}^{(3)}O_{k,m} = {}^{(3)}I_{k,m} \tag{2–27}$$

$$^{(3)}I_{\overline{k,m}} = w_{\varepsilon}^{(k,m)(2)}O_{k,m} + w_{\mathrm{C}}^{(k,m)}, \qquad {}^{(3)}O_{\overline{k,m}} = {}^{(3)}I_{\overline{k,m}} \tag{2–28}$$

$$^{(3)}I'_{\overline{k,m}} = -{}^{(2)}O_{k,m}, \qquad\qquad\qquad {}^{(3)}O'_{\overline{k,m}} = \log {}^{(3)}I'_{\overline{k,m}} \tag{2–29}$$

where $^{(3)}I_{k,m}$ and $^{(3)}O_{k,m}$ denote the input-output of training target classes, and $^{(3)}I_{\overline{k,m}}$, $^{(3)}O_{\overline{k,m}}$, $^{(3)}I'_{\overline{k,m}}$ and $^{(3)}O'_{\overline{k,m}}$ represent the input-outputs of unlearned classes. The outputs of the third layer represent the log-likelihood of each component based on (2–16) and (2–17), and can be used to calculate the *a posteriori*

probability of each component. Hence, the input-output relation of the fourth layer is defined as

$$^{(4)}I_{k,m} = {}^{(3)}O_{k,m} \tag{2–30}$$

$$^{(4)}O_{k,m} = \frac{p(k \neq 0)}{F_{\mathrm{NN}}} \exp\left({}^{(4)}I_{k,m}\right) \tag{2–31}$$

$$^{(4)}I_{\overline{k,m}} = {}^{(3)}O_{\overline{k,m}} + {}^{(3)}O'_{\overline{k,m}} \tag{2–32}$$

$$^{(4)}O_{\overline{k,m}} = \frac{p(k = 0)}{F_{\mathrm{NN}}} \exp\left({}^{(4)}I_{\overline{k,m}}\right) \tag{2–33}$$

$$F_{\mathrm{NN}} = p(k \neq 0) \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} \exp\left({}^{(4)}I_{k',m'}\right)$$
$$+ p(k = 0) \sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} \exp\left({}^{(4)}I_{\overline{k',m'}}\right). \tag{2–34}$$

The fifth layer consists of $K + 1$ units corresponding to the number of target classes, and is used to estimate the *a posteriori* probability of each class $p(k|\boldsymbol{\theta'})$. The input-output relation of this layer is defined as

$$^{(5)}I_k = \begin{cases} \displaystyle\sum_{m'=1}^{M_k} {}^{(4)}O_{k,m'} & (k \neq 0) \\ \displaystyle\sum_{k'=1}^{K} \sum_{m'=1}^{M_{k'}} {}^{(4)}O_{\overline{k',m'}} & (k = 0) \end{cases} \tag{2–35}$$

$$^{(5)}O_k = {}^{(5)}I_k. \tag{2–36}$$

2.3.2 Parameter optimization

The nth training sample consists of the input vector $\boldsymbol{x}_n$ and the teacher vector $\boldsymbol{T}^{(n)} = \left(T_0^{(n)}, \cdots, T_K^{(n)}\right)^{\mathrm{T}}$ $(n = 1, \cdots, N)$. When $\boldsymbol{x}_n$ belongs to class $\hat{k}$, $T_{\hat{k}}^{(n)} = 1$ and $T_k^{(n)} = 0$ for all other classes. If the probability of each class's occurrence is expressed by continuous values, $\sum_{k=0}^{K} T_k^{(n)} = 1$ and $T_k^{(n)} \geq 0$. To minimize Kullback-Leibler divergence between $\boldsymbol{T}^{(n)}$ and network output $^{(5)}O_k$, network
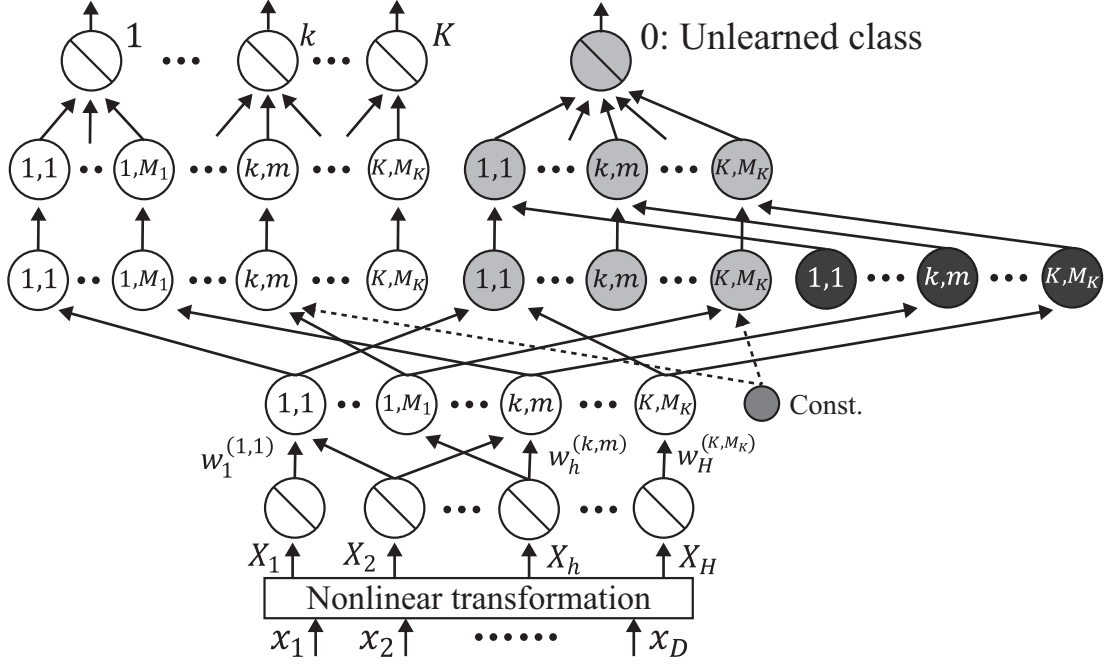
Fig. 2.2: Proposed neural network structure

parameters are trained based on an error function defined as

$$J = \sum_{n=1}^{N} J_n = -\sum_{n=1}^{N} \sum_{k=0}^{K} T_k^{(n)} \log {}^{(5)}O_k^{(n)}. \tag{2--37}$$

The *a priori* probability of the unlearned class $p(k=0)$ is initialized using the appropriate preset value $(p(k=0) = 0.01)$, and the backpropagation algorithm is applied. The weight of $\boldsymbol{w}^{(k,m)}$ is modified both by differentiation of the evaluation value $g_h^{(k,m)}$ and a learning ratio calculated as

$$g_h^{(k,m)} = \sum_{n=1}^{N} \frac{\partial J_n}{\partial w_h^{(k,m)}} \tag{2--38}$$

$$w_{h,\text{New}}^{(k,m)} = w_{h,\text{Old}}^{(k,m)} - \eta_{\boldsymbol{w}} g_h^{(k,m)} \tag{2--39}$$

Here, $\eta_{\boldsymbol{w}} > 0$ is the preset learning ratio for $\boldsymbol{w}^{(k,m)}$ training. $w_{\text{N}}^{(k,m)}$, $w_{\text{C}}^{(k,m)}$ and $w_{\varepsilon}^{(k,m)}$ are also updated using the same algorithm with corresponding learning ratios η_A, η_B and η_E. Differentiation of evaluation values for each weight are

determined using *chain regulation* and expressed by the network input-output relationship

$$
\frac{\partial J_n}{\partial w_h^{(k,m)}} = \left\{ \left({}^{(5)}O_k^{(n)} - T_k^{(n)} \right) \frac{{}^{(4)}O_{k,m}^{(n)}}{{}^{(5)}O_k^{(n)}} + \left(w_\varepsilon^{(k,m)} + \frac{1}{{}^{(2)}O_0^{(n)}} \right) \right.
$$
$$
\left. \times \left({}^{(5)}O_0^{(n)} - T_0^{(n)} \right) \frac{{}^{(4)}O_{k,m}^{(n)}}{{}^{(5)}O_0^{(n)}} \right\} X_h \tag{2–40}
$$

$$
\frac{\partial J_n}{\partial w_{\mathrm{N}}^{(k,m)}} = \left({}^{(5)}O_k^{(n)} - T_k^{(n)} \right) \frac{{}^{(4)}O_{k,m}^{(n)}}{{}^{(5)}O_k^{(n)}} \tag{2–41}
$$

$$
\frac{\partial J_n}{\partial w_{\mathrm{C}}^{(k,m)}} = \left({}^{(5)}O_0^{(n)} - T_0^{(n)} \right) \frac{{}^{(4)}O_{k,m}^{(n)}}{{}^{(5)}O_0^{(n)}} \tag{2–42}
$$

$$
\frac{\partial J_n}{\partial w_\varepsilon^{(k,m)}} = \left({}^{(5)}O_0^{(n)} - T_0^{(n)} \right) \frac{{}^{(4)}O_{k,m}^{(n)} \, {}^{(2)}O_{k,m}^{(n)}}{{}^{(5)}O_0^{(n)}}. \tag{2–43}
$$

However, as $\boldsymbol{w}^{(k,m)}$ values involve weak statistical constraints derived from $\boldsymbol{\Sigma}_{k,m}$ covariance matrices, some weights corresponding to diagonal components of $\boldsymbol{\Sigma}_{k,m}^{-1}$ will be negative. In particular,

$$
w_{H'}^{(k,m)} < 0, \ H' = \left\{ \frac{d}{2} \left(2D + 3 - d \right) \mid d = 1, \cdots, D \right\} \tag{2–44}
$$

must be fulfilled. If this constraint deviates in the update process, an appropriate negative value is reassigned and training is restarted. If ${}^{(3)}I'_{\overline{k,m}} < \rho$, outputs of the unlearned class unit in the second layer are overwritten as ${}^{(3)}I'_{\overline{k,m}} = \rho$, where ρ is a sufficiently small positive constant.

2.3.3 Estimation of statistical model parameters based on trained NACGMN

The proposed NACGMN is a probabilistic neural network incorporating GMM and CGMM parameters, enabling calculation of *a posterior* probabilities in consideration of unlearned class variables with Log-linearization transformation for $\boldsymbol{\theta}$. Thus, while a NACGMN model enables Bayesian discrimination based on

probabilistic modeling, it also displays the properties of an original generative model. However, the recovery of statistical parameters from optimized weight coefficients has not been discussed in relation to conventional methods [5, 22], and trained NACGMN output cannot be simply treated as generative data. The readability of complex neural networks has also long been problematic, and various ways to visualize internal network structures have been proposed [38, 39]. Establishing a statistical parameter extraction method for a NACGMN model will enable mathematical expression of the discrimination space and improve readability. Accordingly, this paper proposes NACGMN parameter extraction employing a gradient-based method.

The covariance matrix $\boldsymbol{\Sigma}_{k,m}$ can be restored from the weight coefficient $\boldsymbol{w}^{(k,m)}$. Based on Eq. (2–13),$w_h^{(k,m)}$ $(h = D + 1, \cdots, H - 1)$ corresponds to each element of the precision matrix $\boldsymbol{\Lambda}_{k,m} \in \Re^{D \times D} \equiv \boldsymbol{\Sigma}_{k,m}^{-1}$. Hence, $\hat{\boldsymbol{\Lambda}}_{k,m}$ estimated from a NACGMN model can be derived as

$$\hat{\boldsymbol{\lambda}}_{k,m} = \begin{bmatrix} w_{D+1}^{(k,m)} & w_{D+2}^{(k,m)} & \cdots & w_{2D}^{(k,m)} \\ 0 & w_{2D+1}^{(k,m)} & \cdots & w_{3D-1}^{(k,m)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{H-1}^{(k,m)} \end{bmatrix} \tag{2–45}$$

$$\hat{\boldsymbol{\Lambda}}_{k,m} = -\hat{\boldsymbol{\lambda}}_{k,m} - \hat{\boldsymbol{\lambda}}_{k,m}^{\mathrm{T}}. \tag{2–46}$$

Here, $\hat{\boldsymbol{\lambda}}_{k,m} \in \Re^{D \times D}$ is an upper triangular matrix. From these results, the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_{k,m} \in \Re^{D \times D}$ can be calculated using

$$\hat{\boldsymbol{\Sigma}}_{k,m} = \hat{\boldsymbol{\Lambda}}_{k,m}^{-1}. \tag{2–47}$$

The mean vector $\hat{\boldsymbol{\mu}}_{k,m} \in \Re^D$ can then be estimated from gradient descent using $\hat{\boldsymbol{\Lambda}}_{k,m}$. As $q(\boldsymbol{x}; k, m)$ is a convex upward multivariate quadratic function, the point with the maximum value coincides with $\hat{\boldsymbol{\mu}}_{k,m}$, and gradient descent can be applied to $\hat{\boldsymbol{\mu}}_{k,m}$ determination. An arbitrary initial value is set for $\hat{\boldsymbol{\mu}}_{k,m}$, which is reiterated until a preset threshold value is satisfied. The i-th update formula is

expressed as

$$\hat{\boldsymbol{\mu}}_{k,m}^{(i+1)} = \hat{\boldsymbol{\mu}}_{k,m}^{(i)} + \eta_{\hat{\boldsymbol{\mu}}} \nabla q(\hat{\boldsymbol{\mu}}_{k,m}^{(i)}; k, m) \qquad (2\text{–}48)$$

$$\nabla q(\hat{\boldsymbol{\mu}}_{k,m}^{(i)}; k, m) = \boldsymbol{W} - \hat{\boldsymbol{\Lambda}}_{k,m} \hat{\boldsymbol{\mu}}_{k,m}^{(i)} \qquad (2\text{–}49)$$

$$\boldsymbol{W} = \left[w_1^{(k,m)}, w_2^{(k,m)}, \ldots, w_D^{(k,m)} \right]^{\mathrm{T}} \qquad (2\text{–}50)$$

Here, $\boldsymbol{W} \in \Re^D$ is a matrix with elements extracted from $\boldsymbol{w}^{(k,m)}$. The absolute value of $\nabla q(\hat{\boldsymbol{\mu}}_{k,m}^{(i)}; k, m)$ can be used as a threshold, and iteration ends when the differential value is sufficiently small. The mixture coefficients $\hat{\alpha}_{k,m}$, $\hat{\beta}_{k,m}$ and $\hat{\varepsilon}_{k,m}$ of the relevant CGMMs are then calculated analytically. Values are estimated based on the definition using

$$\hat{\varepsilon}_{k,m} = \frac{1}{w_\varepsilon^{(k,m)}} \qquad (2\text{–}51)$$

$$\hat{\alpha}_{k,m} = \exp\left(w_{\mathrm{N}}^{(k,m)} + \frac{D}{2} \log 2\pi + \frac{1}{2} \log \left| \hat{\boldsymbol{\Sigma}}_{k,m} \right| \right) \qquad (2\text{–}52)$$

$$\hat{\beta}_{k,m} = \exp\left\{ w_{\mathrm{C}}^{(k,m)} + \frac{D}{2} \log 2\pi + \frac{1}{2} \log \left| \hat{\boldsymbol{\Sigma}}_{k,m} \right| \right.$$
$$\left. + \left(1 + \frac{D}{2} \right) \log \hat{\varepsilon}_{k,m} - \log 2 \right\}. \qquad (2\text{–}53)$$

As estimated mixture coefficients may deviate from the statistical constraint expressed by Eq. (2–9), normalization is performed using

$$\hat{\alpha}_{k,m}^{(\mathrm{Norm})} = \sigma_{\mathrm{Norm}}^{-1} \hat{\alpha}_{k,m} \qquad (2\text{–}54)$$

$$\hat{\beta}_{k,m}^{(\mathrm{Norm})} = \sigma_{\mathrm{Norm}}^{-1} \hat{\beta}_{k,m} \qquad (2\text{–}55)$$

$$\sigma_{\mathrm{Norm}} = \sum_{k=1}^{K} \sum_{m=1}^{M_k} \hat{\alpha}_{k,m} + \sum_{k=1}^{K} \sum_{m=1}^{M_k} \hat{\beta}_{k,m} \qquad (2\text{–}56)$$

Based on the above, all model parameters of the GMM and CGMM for $\boldsymbol{\theta}$ can be extracted from the weight coefficient of NACGMN $\boldsymbol{\theta}'$.

Finally, by examining the rank of $\hat{\boldsymbol{\Sigma}}_{k,m}$, qualitative evaluation of whether the application of NACGMN was appropriate and whether the hyperparameter settings of NACGMN were correct can be performed. In the case where data with a

very complicated distribution structure is the target of classification, the appropriate number of GMM components and the effectiveness of NACGMN cannot be known in advance. If the application of NACGMN is inappropriate, the traininga will not be successful and $\boldsymbol{\theta}'$, in particular $\hat{\boldsymbol{\Sigma}}_{k,m}$, that satisfies the statistical constraints cannot be obtained. Accordingly, the effectiveness of NACGMN can be evaluated by the calculation of eigenvalues of $\hat{\boldsymbol{\Sigma}}_{k,m}$. When $\hat{\boldsymbol{\Sigma}}_{k,m}$ contains a negative eigenvalue, the remediation of parameter settings and the dimension reduction of input data will be required. In contrast, if all the components do not have negative eigenvalues, the application of NACGMN can be found to be appropriate.

2.4 Experiments

2.4.1 Classification performance evaluation with GMM-based data

To evaluate performance with the proposed NACGMN, pattern classification experiments were conducted with two-dimensional artificial data including four training classes. To generate training samples, GMM application with five components was utilized. Five situations with differing numbers of training data were established with training data N_k numbers of 10, 20, 30, 40, 50 and 100 with the total N as $\sum_{k=1}^{3} N_k$. The test data were divided into five classes in consideration of unlearned classes, with each class containing 10,000 samples and information not including unlearned-class data (for a total of 40,000 data points). To evaluate the variance of identification accuracy, ten data sets output from the same GMMs were produced. Scaling to make the maximum training data value 1 and the minimum -1 was also applied to both data types in advance. For NACGMN training there were $M_k = 2$ components, the *a prior* probability of the unlearned class was $p(k=0) = 10^{-1}$, and the learning ratios for each weight were $\eta_{\boldsymbol{w}} = 10^{-3}$, $\eta_{A,B} = 10^{-4}$ and $\eta_{\varepsilon} = 10^{-5}$. In initialization for NACGMN weighting coefficients, the k-medoids method was applied to the training data of each class, and the $\boldsymbol{\theta}$

value obtained was converted to the network coefficient $\boldsymbol{\theta}'$.

NACGMN classification results were compared with those from the method proposed by Shima *et al.* [21], an SVM-based multi-class classifier with a single-class SVM, and an approach combining a probabilistic neural network (LLGMN) [5] and the isolation forest (IF) technique [19]. The number of components in Shima's method was matched by the M_k value for a NACGMN model. For the SVM-based classifier, an RBF (radial basis function) kernel approach was used and all training data were defined as positive in the single-class SVM training phase. The IF approach included 100 trees, and the training method proposed by Liu *et al.* was used to determine the threshold from training data.

Figure 2.3 shows changes corresponding to N_k in the classification ratios of each method, detailing averages for ten test data sets and the related standard error. The outcomes indicate that differences in classification performance among the methods tend to decrease as N_k increases. For $N_k = 100$, the proposed NACGMN and SVM-based classifiers produced high performance ($97.56 \pm 0.24\%$ and $97.24 \pm 0.33\%$, respectively). With very limited training data, the classification ratios of Shima's previous method and the SVM were reduced. As the SVM-based classifier produced a narrow decision region strongly fitted to a small number of training samples, it is inferred that performance was degraded due to increased misclassification for the unlearned class. The proposed method showed higher levels of classification performance ($83.35 \pm 1.41\%$) even for $N_k = 10$, and significant differences were observed among a NACGMN model, SVM and Shima approaches.

The LLGMN+IF approach produced the highest discrimination ratio regardless of the number of training data, presumably due to a lack of appropriate threshold values in IF. Figure 2.4 details decision regions corresponding to each N_k, with (left to right) training data used and the decision regions of NACGMN and LLGMN+IF. In (b) and (c), the decision regions of classes 1 to 4 are shown in blue, yellow-green, ocher and yellow, and the areas of unlearned classes are shown in dark blue. The outcomes indicate that classification performance for

the unlearned class was significantly degraded by inappropriate threshold settings. In particular, the decision region of the unlearned class disappears in the neighborhood of training samples for $N_k = 10$. The proposed NACGMN produced stable decision regions regardless of N, indicating that the incorporated GMM expressed training data distribution correctly.

To demonstrate the additional effectiveness of log-linearization, the training time of NACGMN was compared to the MCMC method. Two situations with differing numbers of training data were established with N_k numbers of 100 and 1,000, and the time, until the end condition (Finish 10,000 iteration or No parameter changes) was satisfied, was measured. To evaluate the variation in training time, NACGMN and GMM with MCMC was trained five times. For both NACGMN and GMM, the number of components was set $M_k = 2$. The MCMC method sampled not only statistical parameters such as the mean vector but also the number of GMM components in the training process.

Figure 2.5 shows the training time of each method in each case. The outcomes indicate that the computational effort of the backpropagation and the MCMC can derive $O(N)$ and $O(N^2)$ respectively. Thus, the proposed method can be found to achieve faster learning than an optimization using sampling techniques.

These desirable results may be attributable to the low number of dimensions of input features and the definition of teacher labels. In a NACGMN model with a small input dimension such as $D = 2$, the number of required parameters and the number of statistical constraints to be satisfied also decrease, since a dimension of the second layer weight $w(k, m)$ is determined by D. Accordingly, it can be inferred that stable training using the gradient descent method was achieved.

Besides, Figure 2.4 (b) also shows that the decision region of NACGMN training classes was larger than the distribution of training data, which may be partly attributable to the teacher vector setting method. The $\boldsymbol{T}^{(n)}$ value used in training is expressed by a discrete value (1 or 0), while NACGMN outputs are continuous values. As a NACGMN model clearly outputs unclear decisions during the generation of decision region maps, a wider discrimination space containing ambiguous
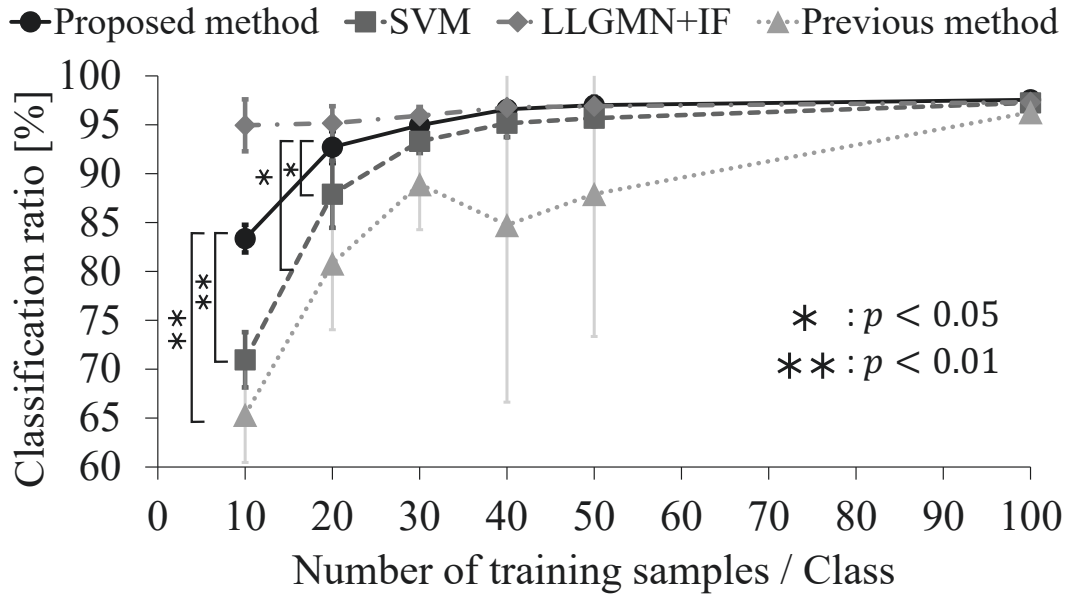
Fig. 2.3: Relationship between training data numbers and classification accuracy

decision regions was produced in all cases. In other words, this phenomenon can be regarded as a brake to prevent overfitting of training data and it is inferred that the phenomenon gave good affection to classification performance in the case that training data is very limited. For more detailed and stable anomaly detection, an appropriate teacher vector based on a posteriori probability is required for each training sample, along with the establishment of a teacher vector generation method.

2.4.2 Evaluation of parameter extraction performance

Performance using the above artificial data approach was evaluated to demonstrate the effectiveness of NACGMN parameter extraction. The NACGMN components and training samples were set as $M_k = 3$ and $N_k = 200$, with the other parameters as detailed in the previous section. After NACGMN training, the proposed parameter extraction was applied and the validity of the restored generative model was examined.

Figure 2.6 (a) shows the training data and original mean vectorss $\boldsymbol{\mu}$ used in
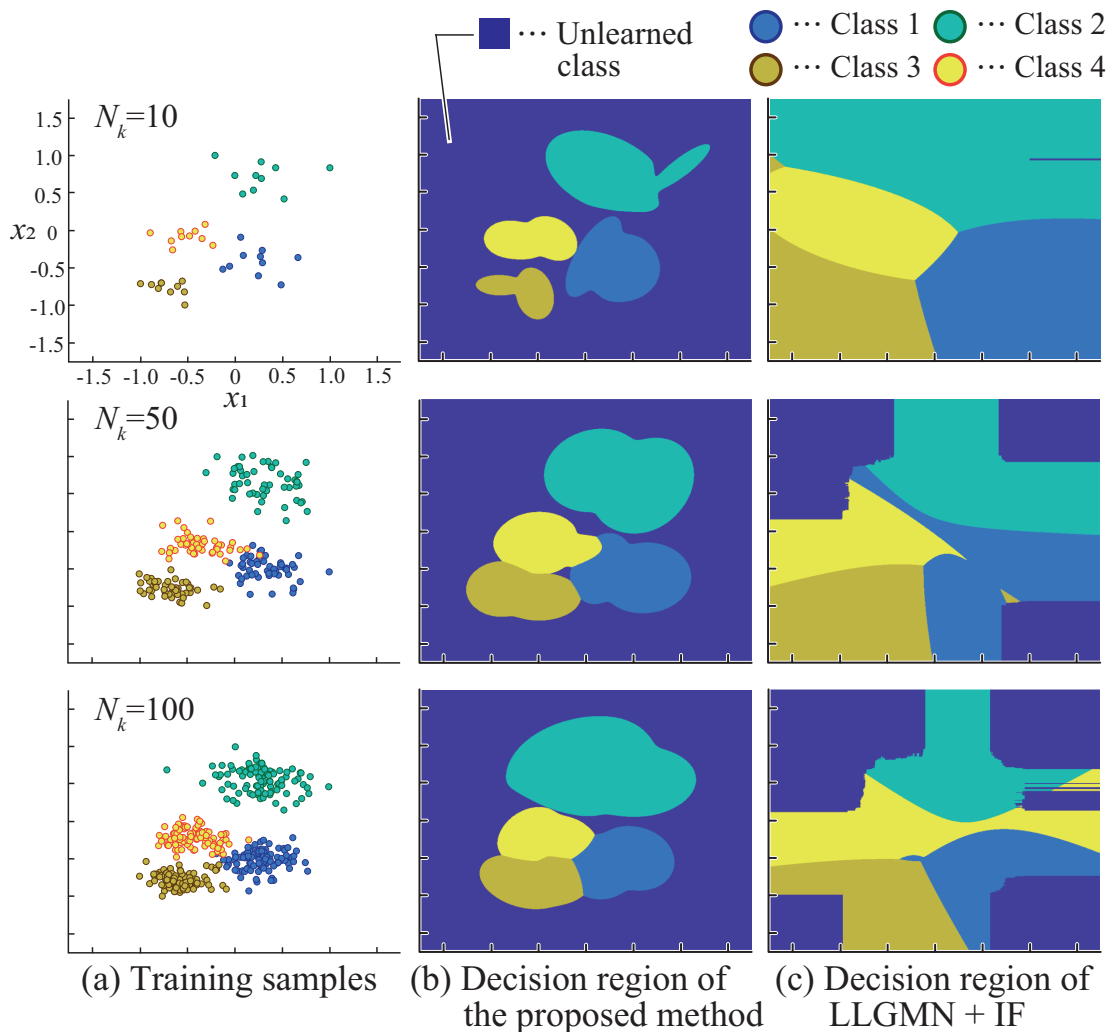
Fig. 2.4: Relationship between training data numbers and decision regions

data generation, and Figure 2.6 (b) shows an NCCGMN model decision region determined with the mean vectors $\hat{\boldsymbol{\mu}}$ extracted. The outcomes indicate that the estimated $\hat{\boldsymbol{\mu}}$ matched the vertices of the Gaussian distributions in the NCCGMN model, suggesting the validity of estimation based on gradient descent. Figure 2.7 (a) shows the decision region produced by the extracted probabilistic model, and Figure 2.7 (b) shows differences in the decision region between the NCCGMN model and the extracted model in black. The results indicate that the shape of the discrimination space for the extracted model was very similar to that of the

Fig. 2.5: Comparison of training time between NACGMN and MCMC

NCCGMN model, demonstrating that the proposed method can be applied to determine this shape roughly and analytically. However, differences are seen, and comparison of Figures 2.6 (b) and 2.7 (a) shows a reduction in the training class decision region. This is attributed to the normalization process used in the estimation of mixtu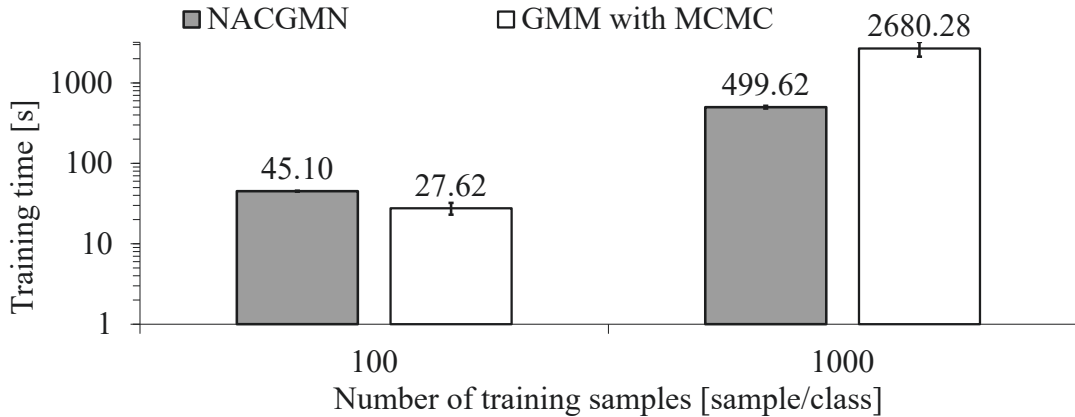re coefficients for the GMM and the CGMM, as log-linearization completely removes mixture coefficient restrictions in NACGMN training, and analytically estimated $\hat{\alpha}$ and $\hat{\beta}$ values may not meet the relevant stochastic constraints. Thus, although estimated mixture coefficients can be converted to appropriate values via normalization, the extraction of GMM and CGMM values fully reproducing those obtained from the NACGMN is problematic. Accordingly, a stochastic constraint related to mixture coefficients must be considered in NACGMN training. In future work, the authors plan to redesign the relevant constraints and employ advanced training algorithms such as AdaGrad [52], which is valid for complex optimization problems.

2.4.3 Classification performance evaluation with non-linearly separable data

To demonstrate the capacity of the NACGMN for non-linearly separable data, evaluation experiments were conducted with two-dimensional data distributed

(a) Training samples (N_k=200)
with original mean vectors

(b) Decision region of NACGMN
with estimated mean vectors
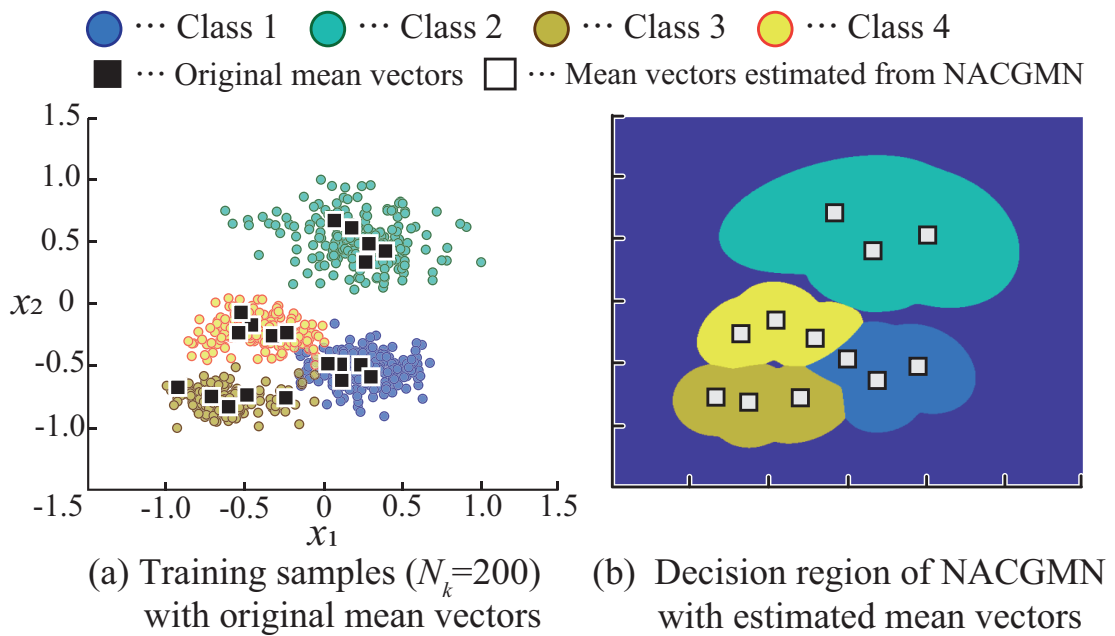
Fig. 2.6: Results of mean vector extraction based on gradient descent

(a) Decision region of extracted
GMMs and CGMMs

(b) Difference between decision
regions of NACGMN and
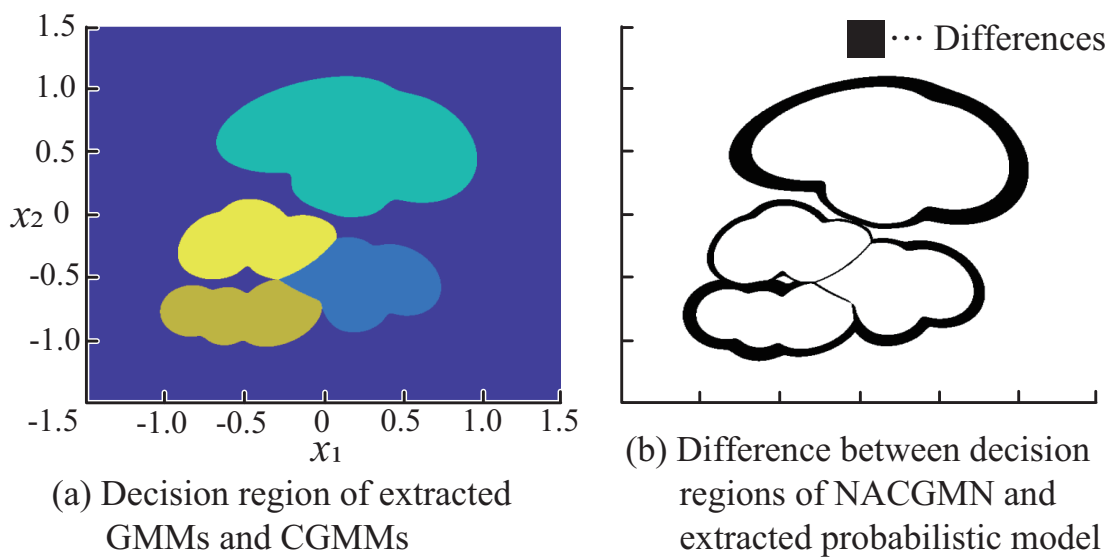extracted probabilistic model

Fig. 2.7: Decision region evaluation for the extracted probabilistic model

concentrically. Figure 2.8 shows generated training samples with four circles, each representing a cluster targeted for training. For trained classes, a total of
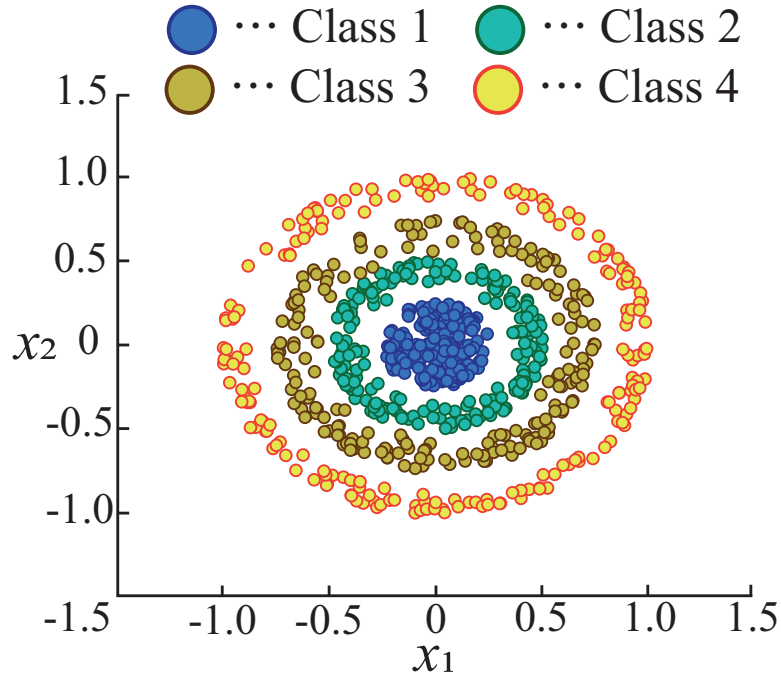
Fig. 2.8: Training samples of non-linearly separable data

800 samples (200 for each class) were used as training data. For evaluation, a total of 40,000 samples (10,000 for each class) with no abnormalities in the unlearned class were employed. To evaluate variations in discrimination accuracy, five sets of training and classification were performed on the same data set. Comparison involved the approach reported Shima [21], an SVM-based classifier approach and a method combining LLGMN and IF. The GMM components of the NACGMN and Shima's approach numbered $M_k = 8$, and an RBF kernel method was employed for SVM-based classification.

Figure 2.9 details decision regions determined with each method. The blue, yellow-green, ocher and yellow areas represent classes 1 to 4, respectively, and the dark-blue area shows the unlearned-class region. The outcomes indicate that the SVM-based classifier determined the most appropriate decision region, with appropriate concentric-circle structure representation. The NACGMN also supported appropriate region decision with increased M_k values, thereby demonstrating NACMN applicability. Although other comparative methods produced the

concentric-circle structure of classes 1 to 3, suitable decision regions for Class 4 and the unlearned class were not achieved. The results of LLGMN+IF may be attributed to a lack of training data and inappropriate thresholding. The results of Shima's approach infer that CGD optimization was inadmissibly difficult due to increased M_k values. Comparison of Figures 2.9 (a) and (d) indicates that NACGMN application sufficiently resolved the limitations of Shima's training method.

Discrimination accuracy from the NACGMN and SVM-based classifier, which produced appropriate decision regions, was also compared. Figure 2.10 shows average classification ratios and standard errors, with "Average" representing the average for all classes. The accuracies of the NACGMN and SVM were $97.50 \pm 1.37\%$ and $99.58 \pm 0.22\%$, respectively, representing no significant difference. Thus, the NACGMN achieved high classification performance comparable to that of the SVM-based classifier. However, its accuracy was lower for classes other than 1, and especially for Class 4. Figure 2.9 (b) shows multiple occasions when the concentric structure was interrupted, which may be attributable to an inappropriate number of components. The representation of concentric circle structures may require more components, but the lower volume of training data for each component hinders optimization. To address this trade-off, the optimum number of components needs to be determined in advance. In future work, the authors plan to introduce non-parametric Bayesian decision with approaches such as the Infinite GMM method [31, 56].

2.4.4 Validity assessment for artificial data with complex distribution

Capacity evaluation experiments were conducted with three-dimensional artificial data to assess the applicability of the NACGMN with data for which GMM representation is unsuitable. Figure 2.11 shows the artificial data employed for the assessment with geometric and complex distributions (from left: Swiss roll, helix and twin peaks). Each data set is divided into four trained classes, with
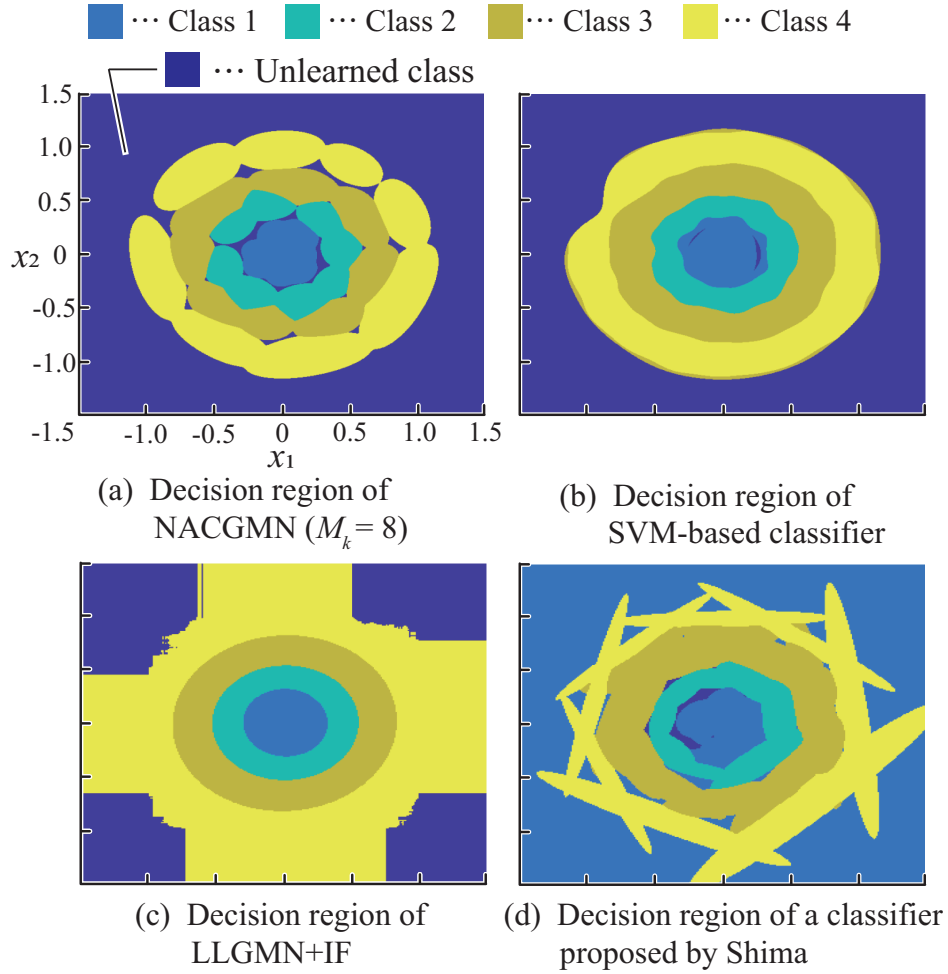
Fig. 2.9: Comparison of decision regions determined with each method

samples of classes 1 to 4 shown in blue, orange, yellow and purple. For trained classes, 400 samples (100 for each class) were used as training data. To evaluate classification accuracy, 40,000 samples (10,000 for each class) with no abnormalities belonging to the unlearned class were used, and training and classification were performed five times with the same data set. A total of $M_k = 6$ GMM components from the NACGMN were used.

Figure 2.12 shows the decision spacing of the NACGMN for each data set. Classified areas for each class are colored, with colorless spaces indicating those classified as unlearned. The outcomes indicate that the discriminant region of the trained classes did not diverge, and the NACGMN was found to produce
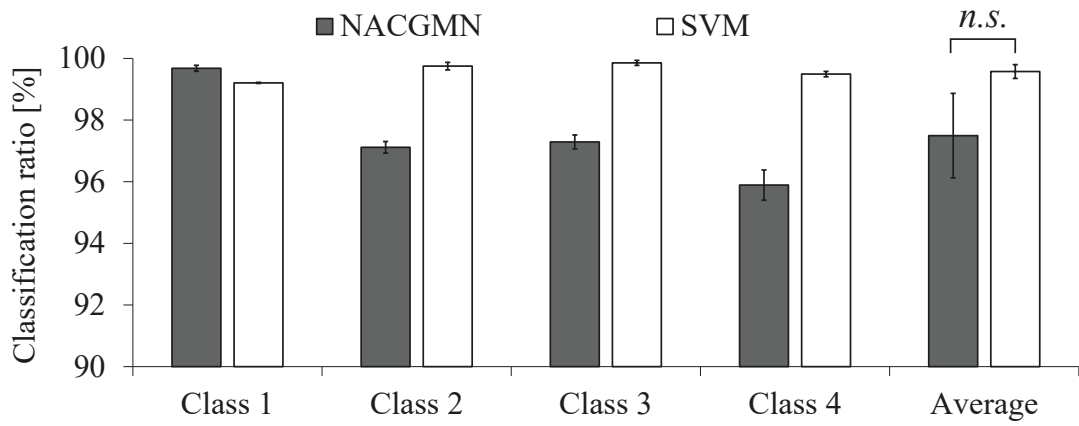
Fig. 2.10: Classification ratios for non-linearly separable circle data

(a) Swiss roll

(b) Helix

(c) Twin-peaks

Fig. 2.11: Complex artificial data types

model parameters that satisfied the statistical constraints from complex geometric data. Figure 2.13 also shows the average classification ratio for trained classes

with the standard error representing the variance of discrimination accuracy associated with initial weights. The results demonstrate high NACGMN classification performance and stable training with small standard deviations for all data.
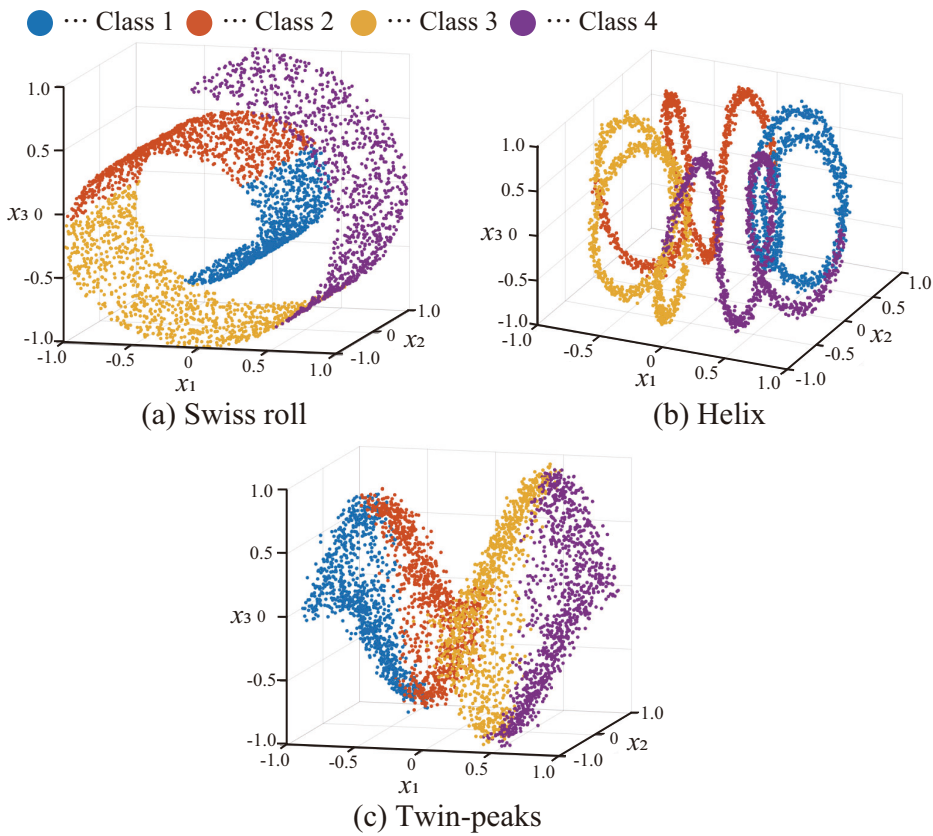
The effectiveness of the proposed method was thus demonstrated, and the suitability of the NACGMN can be evaluated from the determined discrimination space for unlearned classes. Figure 2.14 shows the distribution of training data and the determined decision space in a two-dimensional plane. The upper row represents employed training samples, and the lower row details the decision space of the NACGMN. The Swiss roll data show that the NACGMN has adequate unlearned space inside the roll, and the decision space of the twin-peaks data enabled appropriate representation of related peaks and troughs. However, the NACGMN trained using Helix did not produce an appropriate unlearned class area inside the spiral. This may be attributable to the narrowness of the identification space for the unlearned class inside the spiral and the limitations of semi-supervised learning. As such learning does not involve the use of unlearned-class data, it is not possible to suppress the over-expanded decision space for trained classes. It can be inferred that the narrower the area of the unlearned class surrounded by trained classes is, the higher the probability of disappearance will be. To address this, the use of training data for the unlearned classes and the introduction of penalties for the expansion of decision areas in training target classes should be considered.

2.4.5　Forearm arm motion classification with EMG signals

To verify the validity of the proposed method, experiments involving forearm motion classification were conducted with seven right-handed males. Four electrodes were attached to the right extensor carpi ulnaris (Ch. 1), flexor carpi ulnaris (Ch. 2), abductor pollicis longus (Ch. 3) and extensor carpi radialis (Ch. 4) (Figure 2.15). EMG signals were monitored using wireless EMG loggers (ID3PAD, Oisaka Electronic Equipment) with a sampling frequency of 200 [Hz],

(a) Swiss roll

(b) Helix

(c) Twin-peaks
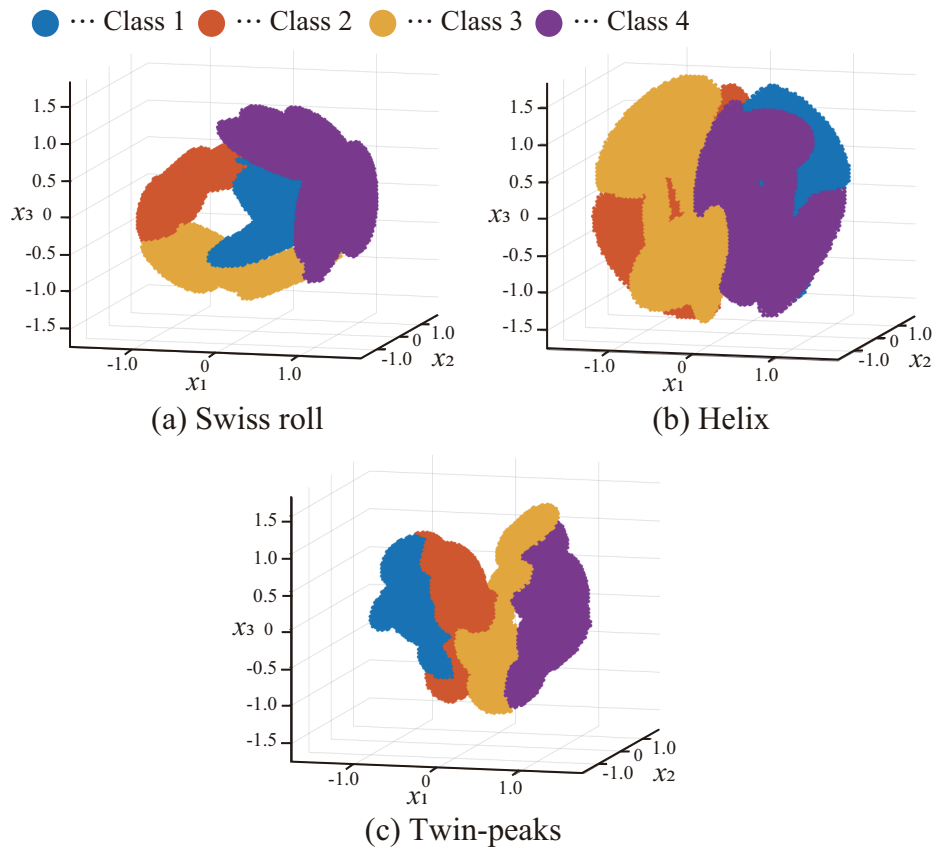
Fig. 2.12: Decision space of NACGMN for each complex data set
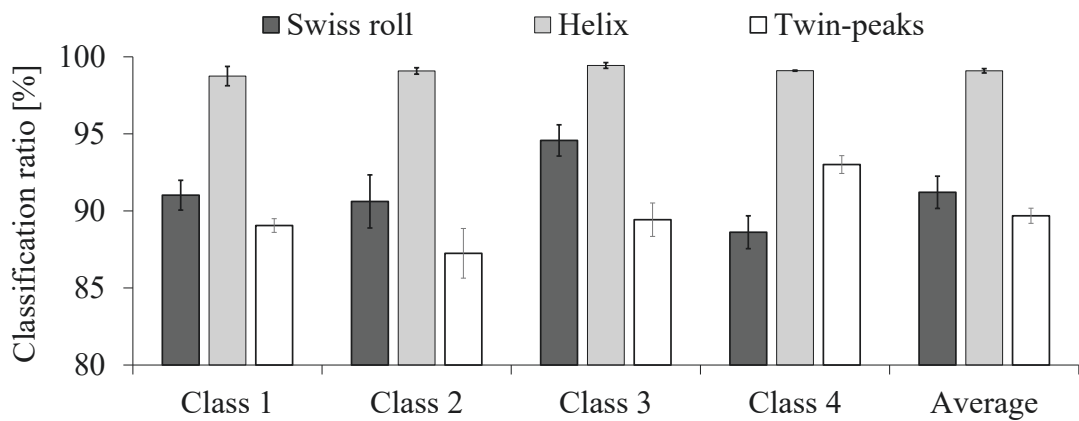
Fig. 2.13: Classification ratios for complex geometric data

and data were collected with the subjects performing (Figure 2.15) four learned motions (wrist flexion: M1; wrist extension: M2; radial flexion: M3; ulnar flexion:
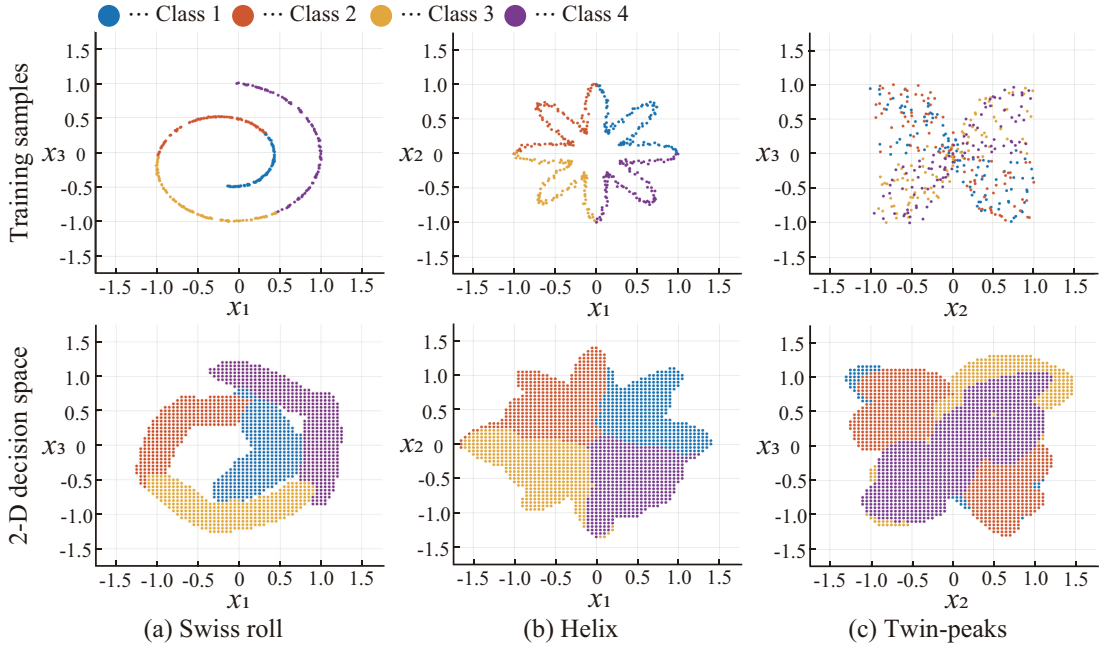
Fig. 2.14: Evaluation of decision spaces

M4) and four unlearned motions (grasp: M5; opening: M6; V-sign: M7; pinching: M8). For the learned class, the subjects performed the motions three times. A total of 1,500 samples for each motion were obtained for use as training data. For data evaluation, the subjects were asked to perform all motions (including those of the unlearned class) twice for approximately three seconds.

The EMG signals were subjected to full-wave rectification to provide feature vectors $\boldsymbol{X}(t) \in \Re^D$ (D: measurement channel count) leveraging past research [4], and smoothing was performed via second-order Butterworth low-pass filtering (cut-off frequency: 2 [Hz]). The smoothed-signal E_d values were normalized using the maximum $E_d^{(\mathrm{max})}$, and

$$E_d^{(\mathrm{norm})}(t) = \frac{E_d(t) - E_d^{(\mathrm{st})}}{E_d^{(\mathrm{max})} - E_d^{(\mathrm{st})}} \tag{2–57}$$

was used to set normalized signals $E_d^{(\mathrm{norm})}$ ($E_d^{(\mathrm{st})}$ average resting EMG signals). The value of $E_d^{(\mathrm{norm})}$ was then normalized so that the sum of all channels was 1.
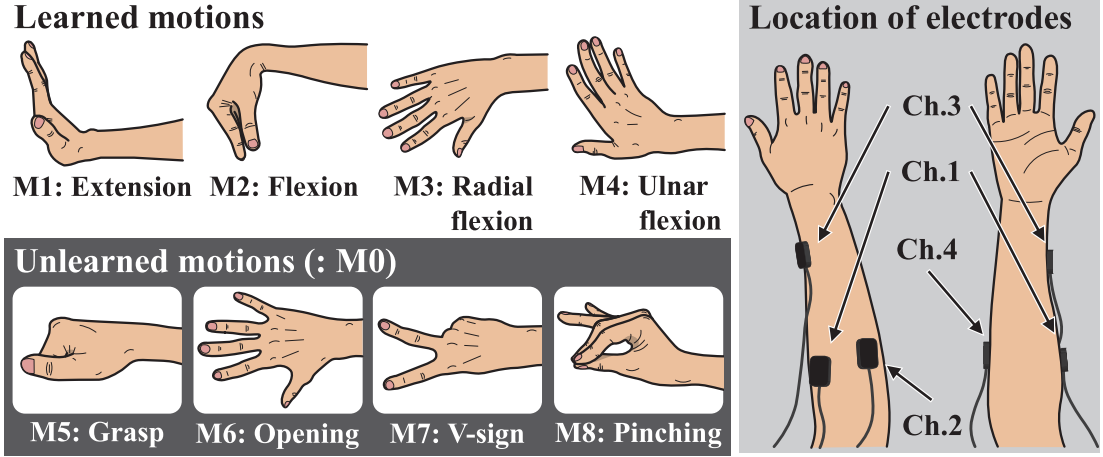
Fig. 2.15: Movements in motion recognition experiments

The feature vector $\boldsymbol{X}(t) = \{\boldsymbol{x}_d(t)\}_{d=1,\ldots,D}$ was represented by

$$x_d(t) = \frac{E_d^{(\text{norm})}(t)}{F(t)} \ (d = 1, \ldots, D) \tag{2–58}$$

$$F(t) = \sum_{d'=1}^{D} E_{d'}^{(\text{norm})}(t). \tag{2–59}$$

Here, force information $F(t)$ was was applied to establish the period of active movement. To ensure precise evaluation of movement, a force threshold F_{th} was established for all subjects, and values of $F(t) > F_{th}$ were taken to represent motion.

The feature vectors $\boldsymbol{x}(t)$ were input to the proposed NACGMN to classify forearm motion. A value of $M_k = 2$ was set based on the results of preliminary experiments, and other parameters were as per the above artificial data experiment. The outcomes of the approach reported by Shima *et al.* [21], SVM-based classifier application [17] and the LLGMN+IF method [5, 19] were compared to evaluate the capacity of the proposed NACGMN. The SVM-based classifier employed a combination of kernels with the highest classification accuracy in preliminary experiments, and the multiclass classifier and single-class SVM employed third-order polynomial and RBF kernels, respectively.

Figure 2.16 details outcomes from classification for the eight forearm motions

(from the top: untreated EMG data, force data from EMG-pattern estimation, and classification with the proposed approach). The grey parts represent periods in which force was below F_{th} (the threshold set in advance), and M0 indicates classification of performed motions as unlearned. The outcomes indicate the capacity of the proposed approach in discrimination for learned and unlearned classes with a low misclassification ratio. The reduced classification ratios were caused by vague motion information in EMG signals. Thus, in periods of motion transition, the likelihood of training classes will be lower. As a result, the *a posteriori* probability of unlearned classes sees a relative increase, and EMG signals may be misclassified. This means that classification with the proposed method can prevent ambiguous results in an unlearned state.

Figure 2.17 shows average classification ratios for each subject and standard errors. "Average" here indicates the average classification ratio for all subjects. Subject C, who was less experienced in the use of EMG signals to classify movements, had the lowest ratios. Higher discrimination ratios are observed for the other subjects, who were accustomed to motion classification experiments involving the use of EMG signals. The lower classification capacity may be attributable to increased variations in EMG patterns. Due to Subject C's lack of experience, appropriate pre-experiment training for improved performance was considered necessary. However, the proposed method produced remarkable classification results even with this level of inexperience, demonstrating its superiority over other anomaly detection approaches. Average classification ratios for all subjects also exhibited significant variations.

Figure 2.18 details average classification ratios and standard errors between learned and unlearned classes for all subjects. The proposed NACGMN exhibited the highest discrimination rate for trained classes, with a classification ratio of $91.97 \pm 2.78\%$. In contrast, the classification ratio of the SVM classifier was significantly reduced for learned classes. This may be attributable to a lack of training data, and it is inferred that single-class SVMs did not obtain sufficient discrimination space for trained classes. As a result, many samples belonging

to trained classes were misclassified as unlearned, and the accuracy of unlearned class detection was relatively high. The capacities of other comparative methods were also reduced for both classes, which may be attributable to difficulties in setting appropriate thresholds and in CGD training. The proposed NACGMN and training method are effective in resolving these problems. However, the classification ratio for the NACGMN model unlearned class was $85.35 \pm 10.29\%$ with a large standard deviation. These outcomes indicate that the decision region of the unlearned class changed with each training instance. Accordingly, learning-rule improvement is needed to enable stable training.

Figure 2.19 shows changes corresponding to the number of used training samples in average classification ratios of each method with standard errors. The results empirically confirm that the proposed method stably outperforms alternative approaches with smaller variances even when training data is limited. Although the accuracy tends to decrease as the number of utilized training motions decreases in all methods as with the results of the experiments of artificial data, the proposed method can maintain the identification accuracy of 80% or more with training data including only one-trial. However, considering the implementation of NACGMN in myoelectric prosthetic hands, accurate identification with a classification ratio of over 90% is required. Thus, if there are no restrictions on the measurement time or measurement instruments, it is desirable to collect EMG signals three times or more for each classified motion.

2.5 Concluding remarks

Chapter 2 describes the proposed probabilistic neural network referred as NACGMN (normal and complementary Gaussian mixture network) capable of determining *a posterior* probabilities for unlearned classes. The network incorporates the GMM and CGMM, and enables determination of related statistical model parameters for network weighting. It can be optimized using a backpropagation algorithm and be applied for semi-supervised learning in which only normal
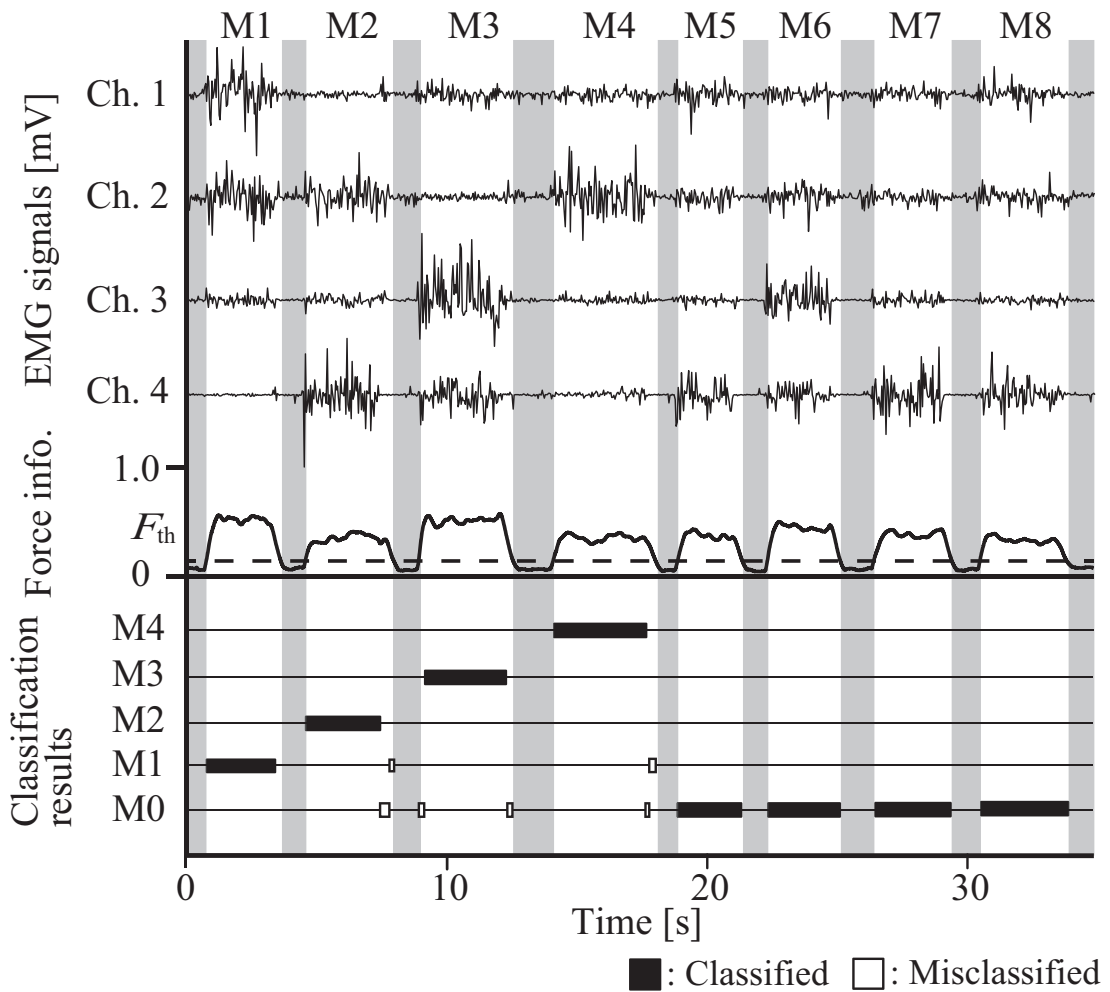
Fig. 2.16: EMGs and classification results

training data without abnormal samples are applied. No trial-and-error determination of thresholds is required, and the approach allows multi-class classification and anomaly detection with a single network. A parameter extraction method for the NACGMN model was also developed, and efforts were made to improve readability.

Two pattern recognition experiments were performed to validate the performance of the approach. In artificial data classification, the NACGMN model was found to enable superior classification even with a small number of training samples. The validity of the proposed parameter extraction method was also
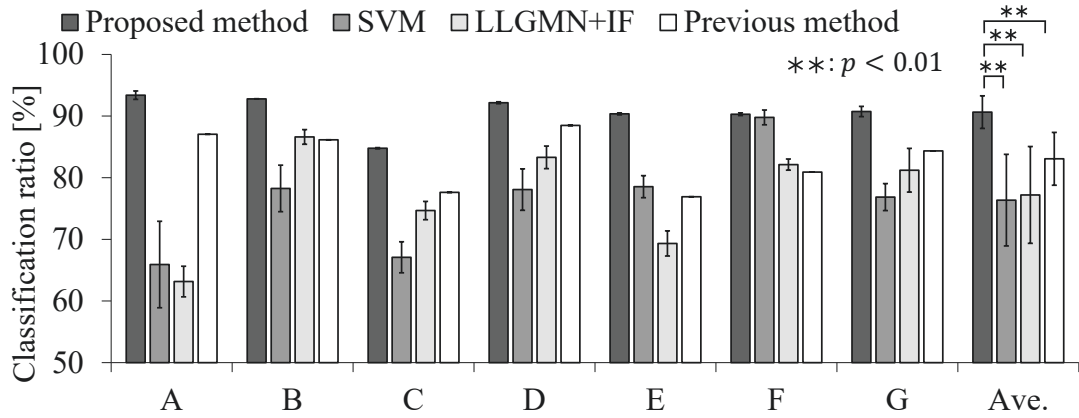
Fig. 2.17: Ratios for unlearned class determination
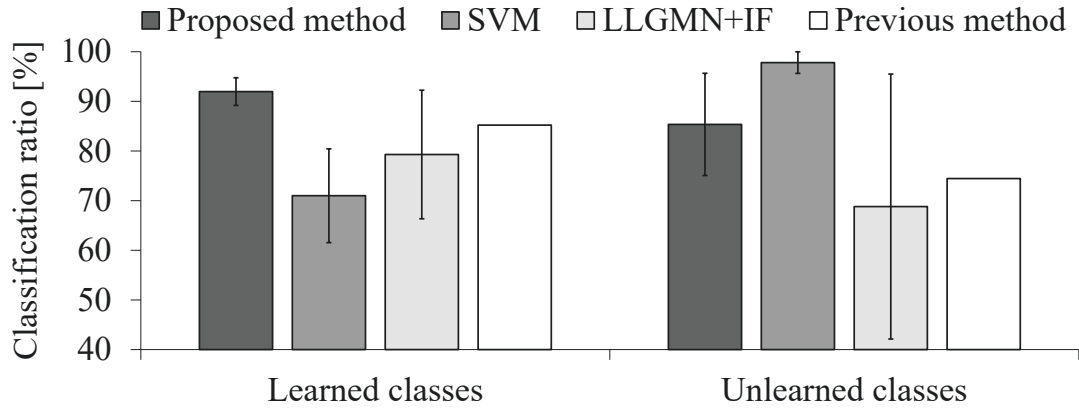
Fig. 2.18: Ratios for learned and unlearned classes

confirmed from the decision region of estimated probability models. In motion classification experiments, the approach was applied for recognition of EMG signals from seven subjects each performing eight forearm motions (four learned and four unlearned). The classification ratios were 91.97%±2.78 (learned) and 85.35%±10.29 (unlearned), representing significantly higher accuracy than other methods for all subjects. Moreover, it is empirically validated that NACGMN stably derives high classification accuracies with a limited number of training data. These outcomes indicate that the proposed approach is suitable for discrimination problems with fewer training samples.
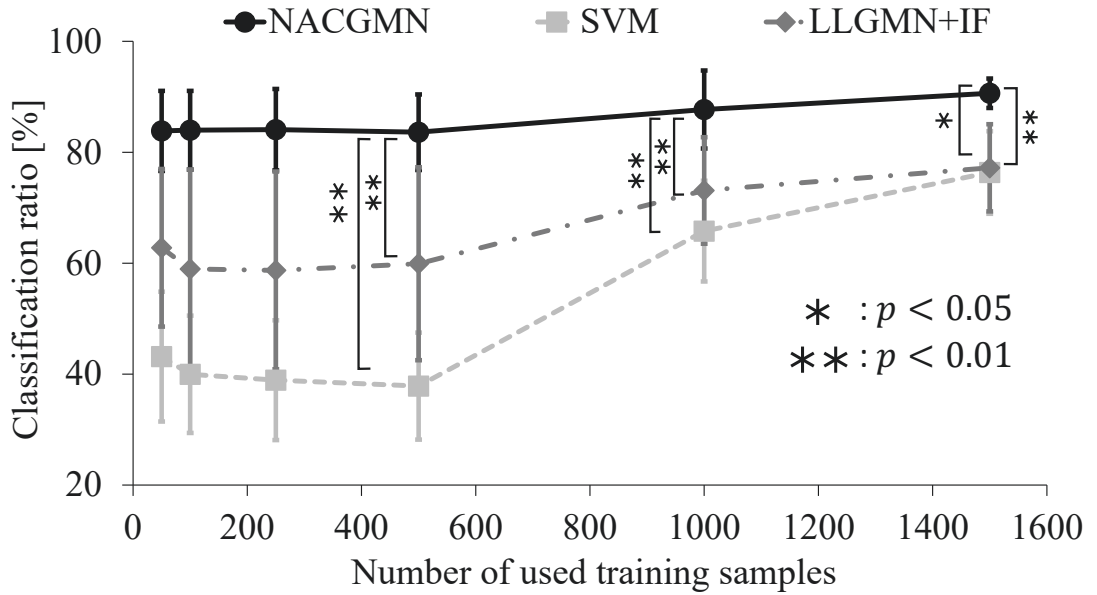
Fig. 2.19: Relationship between training motion numbers and classification accuracy

However, when NACGMN is applied to the classification of unexpected data that does not follow GMM, model parameter setting with trial and error are required, and it is difficult to judge whether the use of NACGMN is appropriate. To deal with this problem, it is desirable to enable to automatically optimize up to model parameters based on nonparametric Bayesian estimation. Although the Bayesian estimation requires a lot of learning data, and when it is difficult to apply it, eigenvalues of $\hat{\mathbf{\Sigma}}_{k,m}$ can be useful to evaluate the applicability of NACGMN. As the presence of negative eigenvalues indicates that the stochastic model contained in NACGMN has collapsed, the problems of parameter setting and feature extraction method can be clarified. In future work, the authors plan to enhance NACGMN functionality, improve learning rules in consideration of statistical constraints, conduct additional performance evaluation, analyze the internal structure of unlearned classes and extract new clusters.

Chapter 3

The one-vs-rest hidden Markov model-based pattern discrimination method with anomaly identification

3.1 Introduction

Chapter 3 proposes a novel one-vs-rest hidden Markov model (OVRHMM) that is applicable to classification of time-series data for both learned and unlearned classes using a hidden Markov model to model the data structures of both classes based on Shima *et al.*'s Gaussian mixture model [21, 37].

In this chapter, Section 3.2 outlines the structure of the proposed hidden Markov model-based classifier, the incorporated probability function for unlearned classes, and the learning methods involved, Section 3.3 explains performance evaluation using artificial time-series data, Section 3.4 discusses the electromyogram discrimination approach and the validity of the method, Section 3.5 covers performance for anomaly identification with neurodegenerative disorders, and Section 3.6 draws conclusions and outlines future study plans.

3.2 Sequential pattern recognition based on one-vs-rest hidden Markov models

3.2.1 Bayesian estimation based on a hidden Markov model

Classification of the signal sequence X into C classes is based on the premise that events (classes) do not arise at the same time. The relevant hidden Markov model is taken as incorporating signal sources with an arbitrary probability density function linked by a primary Markov chain as a variant of the finite state machine. For $\boldsymbol{X} = \{\boldsymbol{x}_n \in \Re^D\}_{n=1,\ldots,N}$ (the observation variable) the following is used to express the probability function for the hidden Markov models of individual classes $p(X|\theta_c)$:

$$p(\boldsymbol{X}|\theta_c) = \sum_{k'=1}^{K_c} \alpha_{c,k'}(N) \tag{3-1}$$

$$\alpha_{c,k}(1) = \pi_{c,k}p(\boldsymbol{x}_1|\phi_{c,k}) \tag{3-2}$$

$$\alpha_{c,k}(n) = p(\boldsymbol{x}_n|\phi_{c,k}) \sum_{k'=1}^{K_c} \alpha_{c,k'}(n-1)A_{c,k',k}. \tag{3-3}$$

Here, $\boldsymbol{\pi}_c = \{\pi_{c,k}\}_{k=1,\ldots,K_c}$ $(c = 1,\ldots,C)$ is the initial state distribution, $\boldsymbol{A}_c = \{A_{c,k',k}\}_{k=1,\ldots,K_c}^{k'=1,\ldots,K_c}$ represents the probabilities of transition from the current state k' to the subsequent k condition, $\boldsymbol{\phi}_c = \{\phi_{c,k}\}_{k=1,\ldots,K_c}$ expresses the variables of output probability distribution for individual states, c represents the number of classes $(c = 1,\ldots,C)$, and K_c expresses the hidden Markov model state count. The hidden Markov model parameters are collectively represented by the expression $\boldsymbol{\theta}_c = \{\boldsymbol{\pi}_c, \boldsymbol{A}_c, \boldsymbol{\phi}_c\}$. In this context, PDFs (Probability density functions) may be included in output probability distribution with expression in the Gaussian mixture model-based hidden Markov model as follows:

$$p(\boldsymbol{x}_n|\phi_{c,k}) = \sum_{m=1}^{M_{c,k}} r_{c,k,m} g\left(\boldsymbol{x}_n; \boldsymbol{\phi}_{c,k,m}\right) \tag{3-4}$$

In this case, the parameters for Gaussian distribution are expressed by $\boldsymbol{\phi}_{c,k,m} = \{\boldsymbol{\mu}^{(c,k,m)}, \boldsymbol{\Sigma}^{(c,k,m)}\}$ $(k = 1,\ldots,K_c, m = 1,\ldots,M_{c,k})$, the term m represents the

number of components ($m = 1, \ldots, M_{c,k}$), $M_{c,k}$ expresses the Gaussian component count for each condition, and the mixture coefficient $r_{c,k,m} > 0$, the mean vector $\boldsymbol{\mu}^{(c,k,m)} \in \Re^D$ and the covariance matrix $\boldsymbol{\Sigma}^{(c,k,m)} \in \Re^{D \times D}$ are associated with the individual components $\{c, k, m\}$. Here, the expression $\sum_{m'=1}^{M_{c,k}} r_{c,k,m'} = 1$ is met. Calculation to determine the a *posteriori* probability of the individual classes $p(c|\boldsymbol{X})$ ($c = 1, \ldots, C$) is implemented with the following in consideration of the above and Bayesian estimation:

$$p(c|\boldsymbol{X}) = \frac{p(c)p(\boldsymbol{X}|\boldsymbol{\theta}_c)}{\sum_{c''=1}^{C} p(c'')p(\boldsymbol{X}|\boldsymbol{\theta}_{c''})}. \tag{3–5}$$

Selection of the class with the maximum a *posteriori* probability from Bayesian discrimination allows recognition of patterns based on the proviso that individual events are dependent.

3.2.2 The proposed hidden Markov model with unlearned class probability density function

The work discussed here involved the use of the Gaussian mixture model reported by Shima *et al.*. This model allows consideration for the probability distribution of unlearned classes, which can be determined from time-series information via the incorporation of PDFs for unlearned classes in hidden Markov models. In this approach, one-versus-the-rest classification is adopted to establish whether the relevant data should be within the $\{c, k, m\}$ component. To this end, the PDF of data outside this scope is established as [21]:

$$p(\boldsymbol{x}_n|\boldsymbol{\phi}_{\bar{c},k,m}) = h(\boldsymbol{x}_n; \boldsymbol{\mu}^{(c,k,m)}, \boldsymbol{\Sigma}^{(c,k,m)}, \boldsymbol{\epsilon}_{c,k,m}) \tag{3–6}$$

$$h(\boldsymbol{x}_n; \boldsymbol{\phi}_{\bar{c},k,m}) = (2\pi)^{-\frac{D}{2}} \left| \epsilon_{2,c,k,m} \boldsymbol{\Sigma}^{(c,k,m)} \right|^{-\frac{1}{2}}$$

$$\times \left(\epsilon_{1,c,k,m}^{\frac{D}{2}} - 1 \right)^{-1} \exp\left(\epsilon_{1,c,k,m}^{-1} \epsilon_{2,c,k,m}^{-1} \right)$$

$$\times \left[\exp\left\{ q(\boldsymbol{x}_n; c, k, m) \right\} - \exp\left\{ \epsilon_{1,c,k,m} q(\boldsymbol{x}_n; c, k, m) \right\} \right]. \tag{3–7}$$

In this calculation, $\boldsymbol{\phi}_{\bar{c},k,m} = \{\boldsymbol{\mu}^{(c,k,m)}, \boldsymbol{\Sigma}^{(c,k,m)}, \boldsymbol{\epsilon}_{c,k,m}\}$ ($k = 1, \ldots, K_{\bar{c}}, m = 1, \ldots, M_{\bar{c},k}$) is defined. The expression $\boldsymbol{\epsilon}_{c,k,m} = \{\epsilon_{1,c,k,m}, \epsilon_{2,c,k,m}\}$ is a new representation for

the form of PDFs for unlearned classes. $\epsilon_{1,c,k,m}$ is applied to establish variations in $h(\cdot)$. The expressions $\epsilon_{1,c,k,m} > 1$ and $\epsilon_{2,c,k,m} > 0$ both hold. In a situation where $\epsilon_{1,c,k,m} = 1$,

$$h(\boldsymbol{x}_n; \boldsymbol{\mu}^{(c,k,m)}, \boldsymbol{\Sigma}^{(c,k,m)}, \boldsymbol{\epsilon}_{c,k,m}) = 0 \tag{3–8}$$

is also valid. Input data are taken to fit with these PDFs, and distribution of information in the unlearned class is established around the individual components. Based on the above, the following is used to express the hidden Markov model output probability distribution in consideration of the unlearned class PDFs:

$$p(\boldsymbol{x}_n | \phi_{\bar{c},k}) = \frac{1}{M_{\bar{c},k}} \sum_{m=1}^{M_{\bar{c},k}} h\left(\boldsymbol{x}_n; \boldsymbol{\phi}_{\bar{c},k,m}\right). \tag{3–9}$$

In this case, $\boldsymbol{\phi}_{\bar{c}} = \{\boldsymbol{\phi}_{\bar{c},k}\}_{k=1,\dots,K_{\bar{c}}}$, $\bar{\boldsymbol{\theta}}_c = \{\boldsymbol{\pi}_{\bar{c}}, A_{\bar{c}}, \boldsymbol{\phi}_{\bar{c}}\}$ are defined.

Redefinition of the Bayesian classification is then implemented to enable the sorting of time-series data into $C + 1$ classes (including C learned classes and the unlearned class), and the probability function of the unlearned class $p(\boldsymbol{X}|\boldsymbol{\theta}_0)$ ($c' = 0$) is expressed using

$$\begin{aligned}
p(c'|\boldsymbol{X}) &= \frac{p(c')p(\boldsymbol{X}|\boldsymbol{\theta}_{c'})}{\sum_{c''=1}^{C} p(c'')p(\boldsymbol{X}|\boldsymbol{\theta}_{c''}) + p(0)p(\boldsymbol{X}|\boldsymbol{\theta}_0)} \\
&= \frac{p(c')p(\boldsymbol{X}|\boldsymbol{\theta}_{c'})}{\sum_{c''=0}^{C} p(c'')p(\boldsymbol{X}|\boldsymbol{\theta}_{c''})}.
\end{aligned} \tag{3–10}$$

In this case, the a *priori* probability of the unlearned class is represented by $p(0)$. $C + 1$ discrimination involves the assumption that no non-defined classes fall within the scope of the predefined components determined with C classifiers. Based on this, the expression $p(\boldsymbol{X}|\boldsymbol{\theta}_0)$ can be used to represent the joint probability of $p(\boldsymbol{X}|\bar{\boldsymbol{\theta}}_c)$, and approximation to establish the a priori probability of the unlearned class is conducted using

$$p(\boldsymbol{X}|\boldsymbol{\theta}_0) \simeq \psi \prod_{c=1}^{C} p(\boldsymbol{X}|\bar{\boldsymbol{\theta}}_c). \tag{3–11}$$

Here, $\psi > 0$.

The eventual a *posteriori* probability can be expressed as follows based on the above two expressions:

$$p(c'|X) = \begin{cases} \dfrac{\psi p(c') \prod_{c=1}^{C} p(\boldsymbol{X}|\bar{\boldsymbol{\theta}}_c)}{f(\boldsymbol{X})} & (c' = 0) \\[3mm] \dfrac{p(c')p(\boldsymbol{X}|\boldsymbol{\theta}_{c'})}{f(\boldsymbol{X})} & (otherwise). \end{cases} \qquad (3\text{--}12)$$

Here,

$$f(\boldsymbol{X}) = \sum_{c=1}^{C} p(c)p(\boldsymbol{X}|\boldsymbol{\theta}_c) + \psi p(0) \prod_{c=1}^{C} p(\boldsymbol{X}|\bar{\boldsymbol{\theta}}_c). \qquad (3\text{--}13)$$

As determination of probability for the hidden Markov models $p(\boldsymbol{X}|\boldsymbol{\theta}_c)$, $p(\boldsymbol{X}|\bar{\boldsymbol{\theta}}_c)$ is established via a forward algorithm [40], $C + 1$ classes (including unlearned ones) can be identified with appropriate estimation of $\boldsymbol{\theta}_c = \{\boldsymbol{\pi}_c, \boldsymbol{A}_c, \boldsymbol{\phi}_c\}$ and $\bar{\boldsymbol{\theta}}_c = \{\boldsymbol{\pi}_{\bar{c}}, \boldsymbol{A}_{\bar{c}}, \boldsymbol{\phi}_{\bar{c}}\}$. The proposed approach is valid for static-model classification if the hidden Markov model state count is $K_c = 1$, and the static model is basically as per the classifier proposed by Shima *et al.* [21].

3.2.3   Determination of model parameters in OVRHMM

The training parameters for $\boldsymbol{\theta}_c$ are determined for optimal probability in the measurement signals $\boldsymbol{X}_{c,r} = \{\boldsymbol{x}_{c,r,n} \in \Re^D\}_{n=1,\ldots,N}$ $(c = 1,\ldots,C, r = 1,\ldots,R)$ using

$$\boldsymbol{\theta}_{c\max} = \arg \max\{p(\{\boldsymbol{X}_{c,r}\}_{r=1,\ldots,R}|\boldsymbol{\theta}_c)\}. \qquad (3\text{--}14)$$

In this case, N represents data length and R expresses the training data set count. The work reported here involved application of the Baum-Welch algorithm [28,40] with the hidden Markov model for learned classes, thereby allowing parameter determination with expectation-maximization algorithm. The learning parameters of this model enable multi-class classification for all classes.

Despite these advantages, the Baum-Welch algorithm does not support estimation of $\bar{\boldsymbol{\theta}}_c$ due to a lack of training data sets for unlearned classes. In the
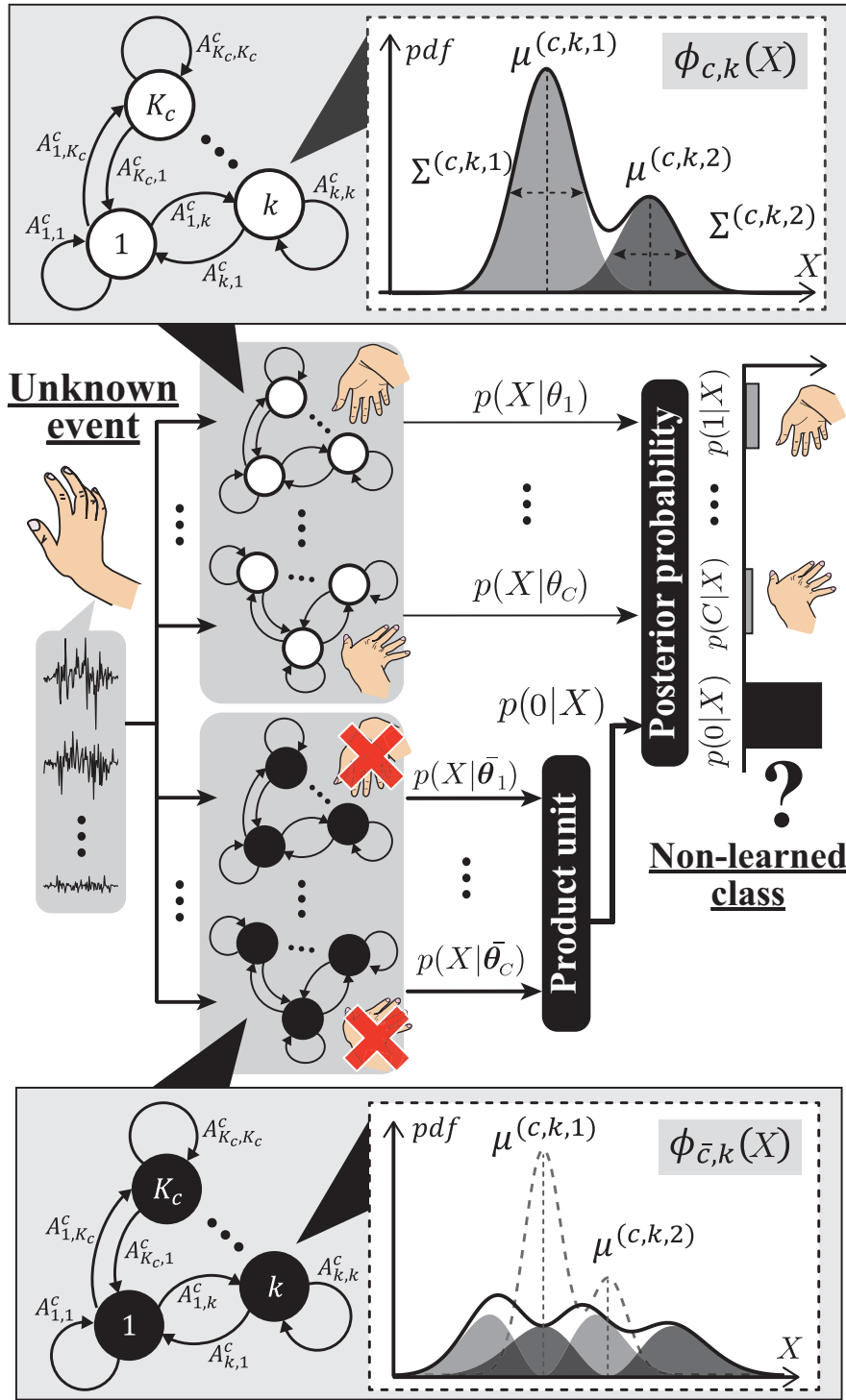
Fig. 3.1: Overview of the proposed OVRHMM-based classifier

hidden Markov model spectrum, models for class c are expressed as H_c and those applied to determine the probability of presence outside this class are expressed as $\overline{H_j}$. Expressions 3–15 and 3–16 below $i = 1, \ldots, C, j = 1, \ldots, C$ are applied to determine the log probabilities $L_{i,j,r}$ and $U_{i,j,r}$ (as obtained from H_c and $\overline{H_j}$ with training data $\boldsymbol{X}_{i,r}$) in one-versus-the-rest classification.

$$L_{i,j,r} = \log[p(\boldsymbol{X}_{i,r}|\boldsymbol{\theta}_j)] \tag{3–15}$$

$$U_{i,j,r} = \log[p(\boldsymbol{X}_{i,r}|\bar{\boldsymbol{\theta}}_j)] \tag{3–16}$$

Based on the premise that $\overline{H_c}$ may be applied to determine the likelihood of values being outside class c, H_c and $\overline{H_c}$ can be deemed incompatible. As a result, $L_{c,c,r} > U_{c,c,r}$ should be valid if $X_{c,r}$ is applied to both hidden Markov models. The expression $L_{\bar{c},c,r} < U_{\bar{c},c,r}$ should also be valid if a vector not belonging to $X_{\bar{c},r}$ ($\bar{c} \neq c$) is applied.

This work involved the assumption that some H_c parameters determined from the Baum-Welch algorithm apply to $\overline{H_c}$ (and in particular to $\boldsymbol{\pi}_{\bar{c}} = \boldsymbol{\pi}_c$ and $\boldsymbol{A}_{\bar{c}} = \boldsymbol{A}_c$) for simplification. Optimization of $\overline{H_c}$ can therefore be achieved via learning of $\boldsymbol{\epsilon}_{c,k,m}$ from learned-class data. A gradient descent approach is used for training on $\boldsymbol{\epsilon}_{c,k,m}$, and the following expression is used to update $\boldsymbol{\epsilon}_{c,k,m}$:

$$\epsilon_{q,c,k,m}^{\text{new}} = \epsilon_{q,c,k,m}^{\text{old}} + \Delta_q \ (\Delta_q > 0, q \in \{1, 2\}). \tag{3–17}$$

In this case, Δ_q ($q = 1, 2$) represents update coefficients. Learning is conducted as follows:

i. Training data $\boldsymbol{X}_{i,r}$ in class i are applied to H_i and $\overline{H_i}$. Comparison of $L_{i,i,r}$ and $U_{i,i,r}$ is then conducted, and Expression (3–17) is applied to update $\epsilon_{2,c,k,m}$ until $L_{i,i,r} > U_{i,i,r}$ ($i = 1, \ldots, C, \ r = 1, \ldots, R$).

ii. Values of $\boldsymbol{X}_{i,r}$ for all classes are input into $\overline{H_i}$, and comparison of the likelihoods thus determined is performed. Expression (3–17) is applied to update $\epsilon_{2,c,k,m}$ until $U_{i,i,r} < U_{j,i,r}$ ($i \neq j, \ i = 1, \ldots, C, \ j = 1, \ldots, C$)

iii. The values of $\boldsymbol{X}_{i,r}$ from 1. are input into $\overline{H_v}$ ($v = 1, \ldots, C$), and comparison of the likelihoods thus determined is performed among values of $\overline{H_v}$. Expression (3–17) is again applied to update $\epsilon_{2,c,k,m}$ until $U_{i,i,r} < U_{i,j,r}$. This procedure produces a value of $\overline{H_c}$ exhibiting levels of likelihood contrary to the value of H_i.

iv. Expression (3–17) is applied to update $\epsilon_{1,c,k,m}$ until $L_{i,i,r} > U_{i,i,r}$ for test data $\tilde{\boldsymbol{X}}_{c,r}$ created using H_c.

$\tilde{X}_{c,r}$ is set as fresh sample content with a level of measurement error approximating probability distribution. Values are produced with based on a primary Markov chain and PDFs with $\boldsymbol{\Sigma}_{\mathrm{Test}}^{(c,k,m)}$ values determined from the $\boldsymbol{\Sigma}^{(c,k,m)}$ of H_c with

$$\boldsymbol{\Sigma}_{\mathrm{Test}}^{(c,k,m)} = \alpha_{c,k,m}\boldsymbol{\Sigma}^{(c,k,m)}. \tag{3–18}$$

In this case, the expression $\alpha_{c,k,m} > 1$ represents a coefficient for increased covariance.

3.3 Experiments

3.3.1 Artificial-data experiments

To evaluate performance with the proposed OVRHMMs, pattern classification experiments were conducted with two-dimensional artificial signals including nine classes (six learned and three unlearned). To generate these signals, hidden Markov models incorporating Gaussian mixture models were utilized. Predefined hidden Markov models can be used to generate complex time-series data based on first-order Markov chains and output PDFs. Each hidden Markov model has two states, and each state has a Gaussian mixture model with two Gaussian distributions. Classification of time-series data generated using existing methods is sub-optimal because data classes overlap in the feature space.
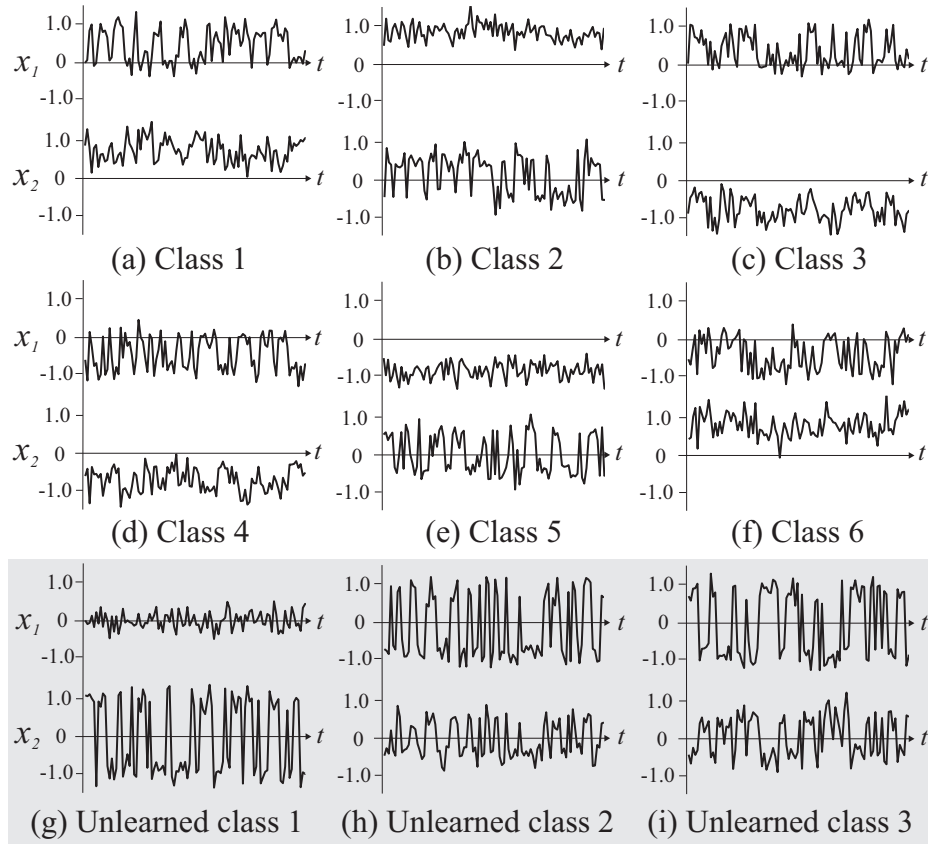
Fig. 3.2: Artificial time-series data generated

For learned classes, a total of 6,000 samples (1,000 for each class) were used as training data. For performance evaluation, three signals in unlearned classes were also generated, and a total of 90,000 samples (learned-class data: 10,000; unlearned-class data: 30,000) were employed. The number of the data samples input to the proposed classifier were 100 for training and 50 for classification. Examples of artificial data are shown in Figure 3.2. To validate the performance of the proposed approach, classification results were compared with those from the method proposed by Shima *et al.* [21] and a multi-class classifier consisting of hidden Markov models with Gaussian mixture models.

The classification ratios obtained with each method for artificial signals are shown in Fig. 3.3. The proposed method produced high ratios (100±0 % for learned classes and 99.17 % for unlearned classes), and those of the previous

Fig. 3.3: Classification results for artificial time-series data

method proposed by Shima *et al.* were 48.82±23.27 % (learned) and 0 % (un-learned). Multi-class classification based on a hidden Markov model also produced high performance for learned classes but was unsuita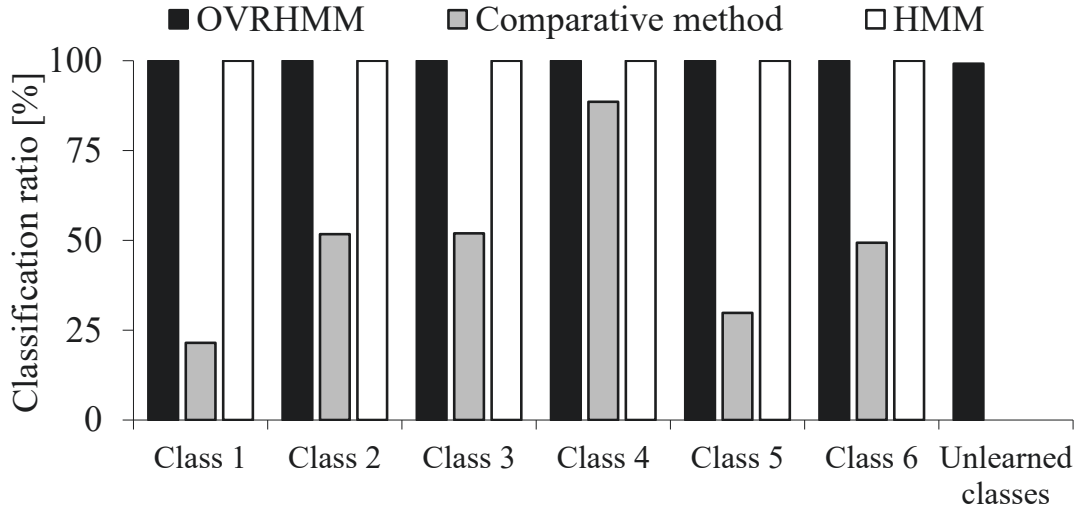ble for unlearned classes. These results demonstrate the proposed method's outstanding classification performance for complex time-series data.

The lower classification ratios may be attributable to the comparative method of Shima *et al.* using only static features of artificial data. As the feature space in the artificial data employed involves class overlap, classification results for overlapping regions can be ambiguous. For example, in Fig. 3.3, the classification ratio for Class 4 is higher than those for Classes 3 and 5. It can be inferred that the decision region for Class 4 was larger than those of the adjacent classes, and data in Classes 3 and 5 were classified as Class 4. As a result, the classification ratio of Class 4 increased and those of Classes 3 and 5 decreased. In addition, as the time-series data of unlearned classes is distributed in learned classes, the previous method resulted in classification of these data as a learned class, and classification performance for unlearned classes deteriorated significantly. However, the as proposed method is applicable to time-axis information, accurate classification can be achieved when feature spaces have overlapping regions.

3.3.2 EMG signal experiments

To verify the validity of the proposed OVRHMM, motion recognition experiments that are the same as Section 2.4.4 were conducted. $\boldsymbol{X}_{c,r}(t) = \{\boldsymbol{x}_d(t)\}_{d=1,\ldots,D}$ (representing the results of feature vector calculation) were input to the proposed hidden Markov models to classify motion. Values of $K_c = 2$, $M_{c,k} = 2$, $\alpha_{c,k,m} = 1.1$, $p(c = 0) = 0.01$ and $\psi = 1$ were applied to the proposed classifier for time-series data application, and comparison of the outcomes with those from NACGMN and OVRGMN reported by Shima *et al.* [21] was performed to evaluate the accuracy of the technique's classification for unlearned class data. This comparison was based on anomaly identification with hidden Markov models and a threshold premised on previously reported trial and error values [42].

Figure 3.4 details outcomes from classification for the eight forearm motions (from the top: untreated EMG data, force data from EMG-pattern estimation, and classification with the approach reported here). The grey parts represent periods in which force was below F_{th} (the threshold set in advance), and M0 indicates classification of performed motions as unlearned. The outcomes indicate the capacity of the approach reported here in discrimination for learned and unlearned classes with a low misclassification ratio. Lower ratios of classification relate to unclear information on movement in EMG signals, meaning that learned-class likelihood will be lower. The *a posteriori* probability of unlearned classes therefore rises in relative terms, and classification of EMG signals may not be accurate. In this context, the proposed method can be implemented to mitigate the instance of ambiguous results with classification for unlearned scenarios.

Figure 3.5 details the subjects' average EMG classification ratios and standard errors. It can be seen that the best performance for learned classes was achieved with the multi-class approach with a hidden Markov model, but this was not applicable to unlearned-class movements. A hidden Markov model with a threshold value performed well in classification of unlearned-class movements, but the challenges inherent in setting the threshold reduced ratios for learned classes considerably. For learned and unlearned classes alike, ratios were higher

with the proposed OVRHMM and OVRGMN proposed by Shima. The incorporation of time information results in the superior performance of the proposed technique with learned classes. Compared with the results of NACGMN, it was shown that OVRHMM achieved equal or better discrimination performance in both classes. As the reason why the performance was not significantly improved, it can be inferred that the covariance matrix could not be obtained correctly due to statistical constraints and the time series characteristics of the EMG signal were not complicated. In the employed EM algorithm, a weighted identity matrix was added to the estimation results in order to realize stable precision matrix calculation. This operation changed the decision space of OVRHMM, and the sufficient improvement was not found to the accuracy of the unlearned class detection.

Average classification ratios and standard errors are shown in Figure 3.6, with "Ave" representing the average for all subjects. Subject C, who was less experienced in the use of EMG signals to classify movements, had the lowest ratios. The lower classification capacity may be attributable to increased variations in EMG patterns. Subject C's familiarity with EMG signal-related interfaces was limited, giving rise to the need for training before the experiment. The subject's relatively positive results despite this inexperience suggest that the method provides more positive performance as compared to other methods for anomaly identification. Average classification ratios for all participants also exhibited significant variations.

3.3.3   Anomaly identification based on gait analysis

The previous section described application of the proposed method to human interfaces and related performance evaluation. The approach can be applied to a variety of anomaly identification problems. This section discusses its potential use in neurological disorder identification using gait data among the range of anomaly identification applications possible.

It is widely known that neurodegenerative conditions such as Parkinson's dis-
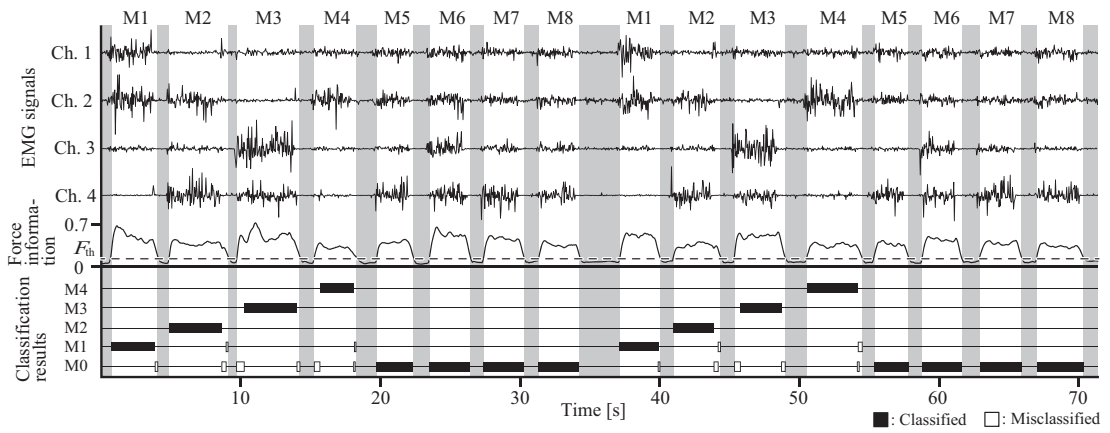
Fig. 3.4: Recorded EMGs and classification results of an OVRHMM-based classifier (Sub. A)
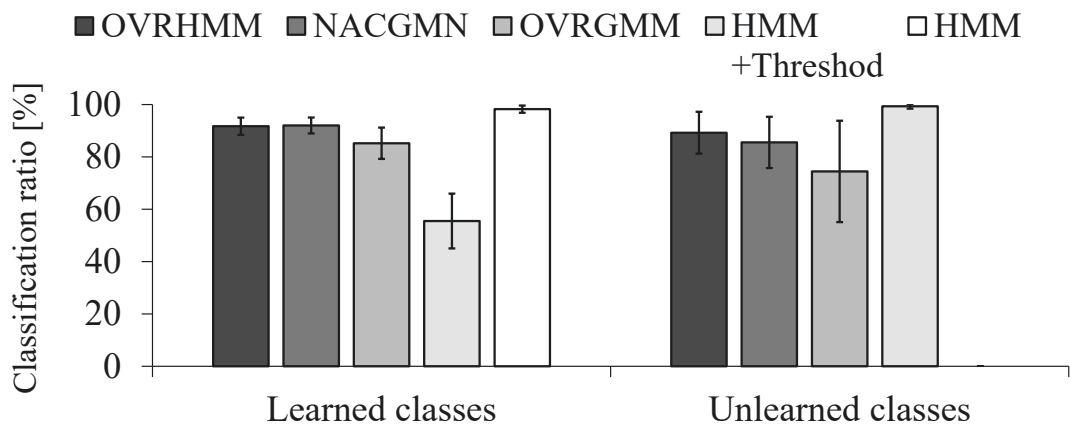
Fig. 3.5: EMG signal classification results

ease and amyotrophic lateral sclerosis (ALS) significantly influence neuromuscular control and motor function [43]. Walking rhythm and kinetics are especially affected, and changes in the foot/ground contact cycle and period are observed [44]. In this regard, analysis of walking characteristics such as stride time allows non-invasive evaluation of such conditions. Extensive research into diagnostic assistance based on gait data has already been performed, and various classification methods have been proposed [36, 41]. However, most such methods require data both on non-impaired subjects and individuals with the relevant condition for classifier training, and have limited application to disorders for which training

Fig. 3.6: Classification ratios for unlearned class detection methods

samples are insufficient. In addition, binary classification between non-impaired subjects and patients with specific conditions involves the problem of misidentification for unlearned diseases. The proposed method addresses these issues by enabling binary classifier training without disorder data. Disorder classes are defined as unlearned, and abnormalities are identified using data exclusively from non-impaired subjects.

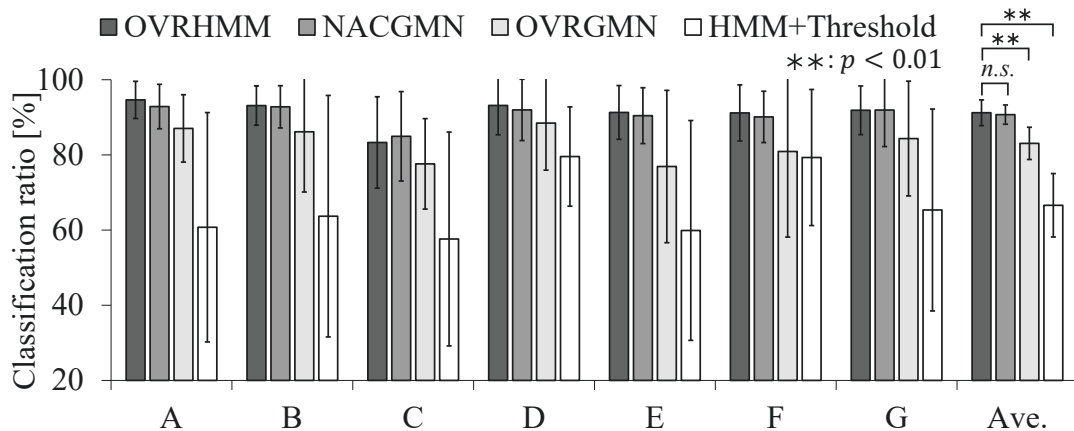Performance evaluation was based on the gait data of Hausdorff *et al.* [43,44], which are available from PhysioNet [45] and relate to 64 subjects with conditions of varying severity (16 non-impaired (age range: 20-74), 15 with Parkinson's (:PD, 44-80), 20 with Huntington's (:HD, 29-71) and 13 with ALS (29-71). The data were recorded using shoe-mounted force sensors, and include stride time, swing time, stance time and double-support periodicity. Figure 3.7 shows stride times for each neurodegenerative condition. Zeng *et al.* proposed a diagnostic system for such conditions based on gait analysis and evaluated related classification performance using this database [41]. In this work, four-dimensional feature vectors were created from data on left/right stance times and leg swing times, and anomaly identification for diagnosis of neurodegenerative conditions was performed using a radial basis function network. Baratin *et al.* proposed a novel feature extraction method based on the discrete wavelet transforms (DWT), and this approach was applied to the same gait database [46]. The paper evaluated the
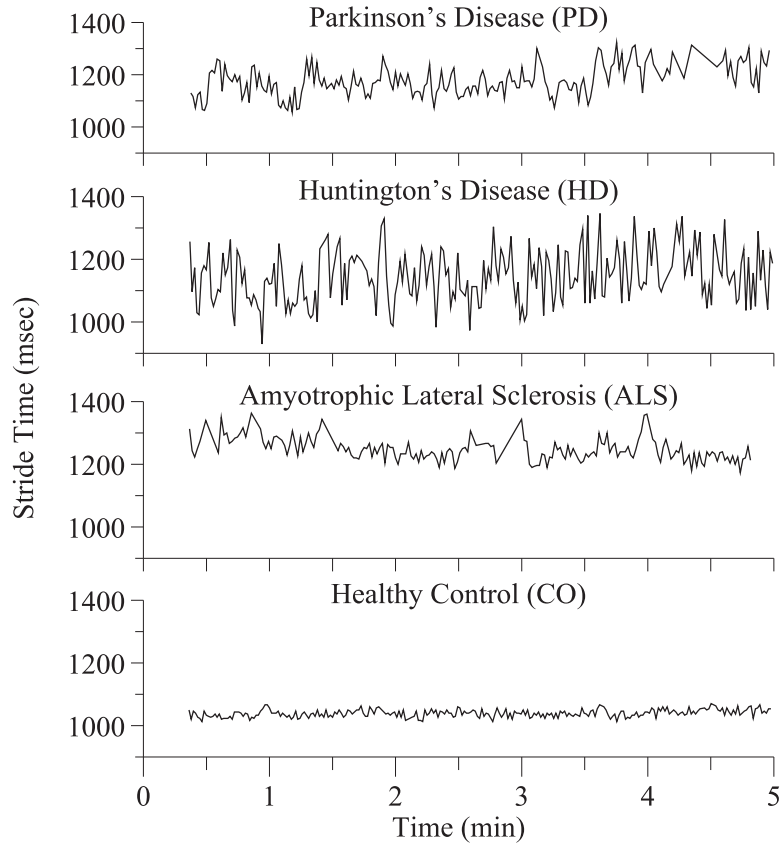
Fig. 3.7: Examples of stride time of each group [43]

classification performance using the linear discriminant analysis (LDA) in various binary classification problems such as Control vs. ALS+PD+HD. Although other methods using the same database have been proposed [47, 48], there are almost no studies that treat four-class classification all classes are training targets. From these studies, the difficulties of multiple disease classification considering healthy peoples were indicated.

The experiments involved the use of the feature vector employed by Zeng *et al.* and a two-dimensional feature vector consisting of swing time and double-support time for the left foot as determined via trial and error from preliminary experiments. These feature vectors are time-series-based, but the sampling interval in the process of feature extraction is inconsistent. To address this problem, cubic spline interpolation was performed for each feature and re-sampling was

iterated with a frequency of 10 [Hz]. The training set consisted of gait data from non-impaired subjects, and the test set consisted of data from all subjects, including those with neurodegenerative conditions. With reference to the method of Zeng *et al.*, the gait data of each subject was divided into a first half used for training and a second half used for evaluation. A total of 1,000 consecutive samples were extracted from the learning data in the training phase, and all test data for each subject were input using the proposed method in the classification experiments. The parameters of the proposed classifier were set as $K_c = 3$, $M_{c,k} = 2$, $\alpha_{c,k,m} = 1.1$, $p(c = 0) = 0.01$ and $\psi = 1$. To validate the capacity of the approach in anomaly identification, the classification results of Zeng *et al.* with the leave-one-out approach are cited [41]. The results of binary classification based on hidden Markov models with training data from both classes were also compared.

Figure 3.8 shows the classification ratios for both classes with the proposed and comparative methods. The proposed method with two-dimensional features exhibits a level of classification performance equal or superior to that of the previous method. It is also noteworthy that the proposed method produced the same classification ratio for the disorder class as the previous method without disorder training data. However, classification ratios under the proposed method with four-dimensional features were lower for both classes. For appropriate classification performance and applicability of the proposed method to diagnostic support, the utilization of appropriate feature vectors needs to be considered.

Although the hidden Markov model approach enabled precise classification for non-impaired subjects, it produced a significantly lower classification ratio for those with disorders. This may be attributable to the use of gait data for all disorder groups in training. As gait data characteristics vary by condition as shown in Fig. 3.7, and the use of information from all subjects resulted in a complex data structure, learning appears to have been insufficient. Although classifiers for individual conditions are needed to help determine whether a subject has a particular condition, the creation of classifiers for a wide variety of impairments is highly

challenging. Additionally, multi-class classification for various conditions will inevitably involve misclassification for certain unlearned instances. The results obtained here indicate the applicability of the proposed method for identifying remarkable characteristics of diseases in subjects. However, as the approach does not necessarily allow recognition of specific conditions, further improvement of classification is required.

To elucidate the classification capacity of the proposed method, the relationship linking classification ratios and condition severity in subjects misidentified as being non-impaired was analyzed. The severity of all patients is expressed in the database via the Hoehn-Yahr scale for Parkinson's disease, the total functional capacity measure (TFCM) for Huntington's disease and the number of months since diagnosis for ALS. Higher Hoehn-Yahr values and lower TFCM values represent greater symptom severity. Averages and standard deviations of severity in the correctly classified group and the misclassified group are shown in Fig. 3.9 (a), (b) indicate minor symptoms in the latter. If symptoms are severe, gait abnormalities facilitate classification. However, the duration of the condition in the misclassified ALS group was longer, as shown in Fig. 3.9 (c). As the standard deviation is considerable for both groups and the prognosis varies among individuals, this cannot be considered an effective index for evaluating discrimination performance with the proposed method. Since ALS severity is generally evaluated from its influence on daily life, evaluation of classification capacity based only on time since diagnosis may be considered ineffective. In future work, the authors plan to develop a learning algorithm that enables anomaly identification in the initial stages of conditions.

3.4 Concluding remarks

This chapter proposes a novel hidden Markov model-based sequential pattern recognition method allowing identification of unexpected unlearned classes in the learning process based on time-series data characteristics. Three pattern recog-
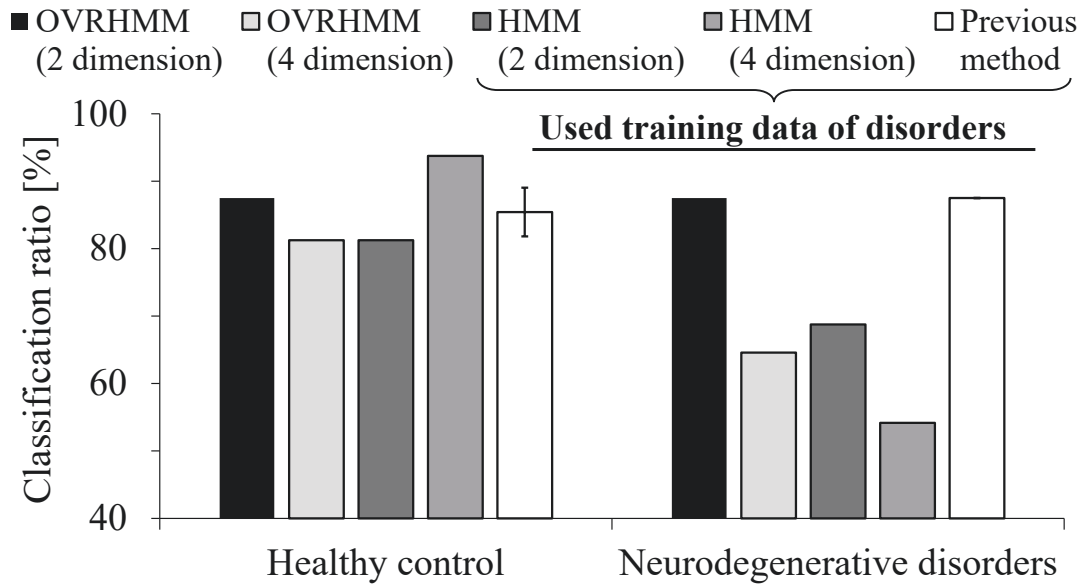
Fig. 3.8: Classification results for gait data of all subjects
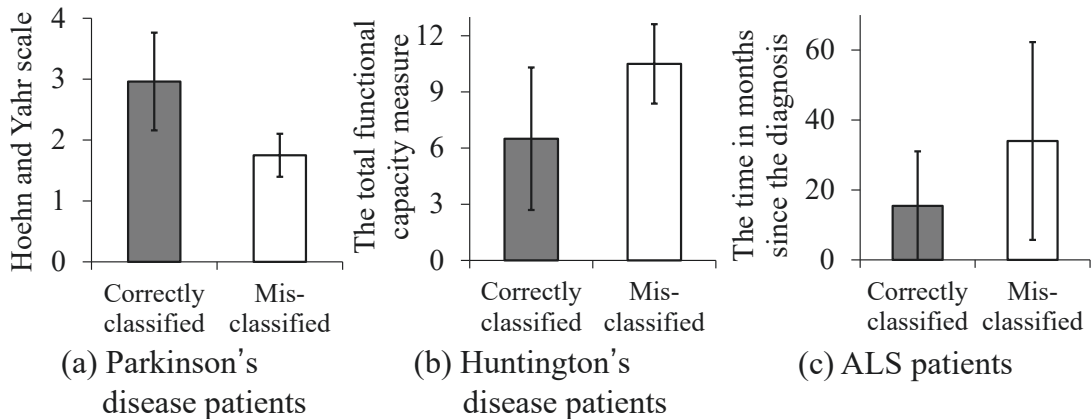
Fig. 3.9: Relationships between degree of severity and misclassification

nition experiments were performed to validate the performance of the approach. In artificial signal classification, complex time-series data were generated from hidden Markov models. The proposed method based on time-series information was found to enable superior classification in anomaly identification. In motion classification experiments, the approach was applied for recognition of EMG signals from seven subjects each performing eight forearm motions (four learned

and four unlearned). The classification ratios were 91.72%±3.30 (learned) and 89.22%±8.01 (unlearned). The results demonstrated that the proposed approach allows classification of unlearned classes with no deterioration in classification skill for learned classes, and that its performance was superior to those of other methods. The approach was also applied to anomaly identification for neurodegenerative conditions using gait data to verify its applicability to professional medical diagnosis. PhysioNet gait data were used, and neurodegenerative-condition patients were placed in the unlearned class for training without disorder gait data. The technique was found to demonstrate a level of performance similar to that of the previous method, in which disorder data are used for training. Based on these characteristics, the applicability of OVRHMM to deal with complicated and special cases with multiple diseases can be expected.

In future work, to establish the effectiveness of OVRHMM, the application of OVRHMM to various time series data needs to be considered. It is inferred that OVRHMM is also effective for temporal data with larger changes and faster transitions than the biological signals employed in this chapter, since the applicability of the proposed method depends on whether the input data can be represented by GMM. Thus, the author aims to expand the range of applications from biological signals and research the applicability of OVRHMM to signals with strong non-stationary. Besides, the authors plan to develop a new unlearned class estimation classifier based on OVRHMM and formulate a new approach to unlearned-class structure analysis. This is expected to support a wide range of applications, including diagnostic assistance and rehabilitation in the medical and welfare fields and abnormality detection in the industrial field. To control myoelectric prosthetic hands [6], the authors also aim to develop a novel hardware implementation method optimized for our approach. Furthermore, to eliminate the complexity of setting model parameters in advance, the introduction of a nonparametric Bayesian optimization method can be considered. However, since those methods required huge training time, the sampling rate of model parameters must be improved. The author plan to log-linearize gauss-Wishart distribution, and the

improvement of calculation speed based on parallel computing.

Chapter 4

Development of anomaly detection with a novel hidden semi-Markov model incorporating unlearned states

4.1 Introduction

This section proposes a novel pattern recognition method based on hidden semi-Markov models (HSMMs) incorporating unlearned states with unpredicted distribution of outliers toward more detailed anomaly detection. The approach incorporates HSMM usage for consideration regarding the time dependency of state transition and efficient training in relation to actual data, and allows anomaly detection at individual time points via estimation of hidden-state transitions inside the model.

Section 4.2 outlines pattern recognition based on the model with unlearned states and parameter estimation, 4.3 covers performance evaluation with artificial signal and application to human behavior classification, and 4.4 draws conclusions and discusses future study plans.

4.2 A hidden semi-Markov model incorporating unlearned states

4.2.1 Estimation for the state transition sequence of a hidden semi-Markov model

The HMM involves a statistical time-series data approach in which signal sources with an arbitrary probability density function are connected via a primary Markov chain. However, it is unsuitable for describing state transitions stochastically with real data, such as speech signals and electrocardiograms (ECGs). The timing of state transition in the handling of biomedical signals depends on the time remaining in each state, making it desirable to utilize an HSMM representing an improved HMM. The HSMM has the state duration density $P_j(d)$ ($j = 1, \cdots, K,\ d = 1, \cdots, D$) representing the probability of maintenance in each state, and allows consideration of the time dependency of state transition. Here, K is the number of HSMM states and D represents the maximum state duration. In the study reported here, an explicit-duration hidden Markov model (ED-HMM) with a relatively simple structure among HSMMs was employed [23, 24]. Given the observed time-series data $\boldsymbol{X} = \{\boldsymbol{x}_t \in \Re^U\}_{t=1,\ldots,T}$, the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ calculated from the ED-HMM is

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{j=1}^{K} \alpha_j(T) \tag{4–1}$$

$$\alpha_j(t) = \sum_{i=1}^{K} \sum_{d=1}^{D} \alpha_i(t-d) a_{i,j} P_j(d) u_t(j,d) \tag{4–2}$$

$$u_t(j,d) = \prod_{\tau=t-d+1}^{t} b_j(\boldsymbol{x}_\tau) \tag{4–3}$$

$$\alpha_j(1) = \pi_j P_j(1) b_j(\boldsymbol{x}_1) \tag{4–4}$$

where π_j ($j = 1, \cdots, K$) is the initial state distribution, $\boldsymbol{A} = \{a_{i,j}\}_{i=1,\ldots,K,\ j=1,\ldots,K}$ represents the probability of transition from the current state i to the next state

j ($a_{i,i} = 0$, $i \in K$) and $b_j(\boldsymbol{x}_t)$ represents the probability of $\boldsymbol{x}_t$ output from state j at time t. Arbitrary PDFs such as the GMM can be incorporated into output probability distribution $b_k(\cdot)$ as with HMMs.

If the set of all HSMM parameters with GMM $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{A}, P_k(d), \boldsymbol{\phi}_k\}_{k=1,\ldots,K}$ is appropriately determined, hidden state transition not directly determined from observation time-series data can be estimated. Based on the use of $\alpha_j(t)$ determined from likelihood calculation, the hidden state sequence $\hat{\boldsymbol{S}} = \{\hat{s}(t)\}_{t=1,\ldots,T}$ can be predicted as

$$\hat{s}(t) = \arg \max_{j \in K} \alpha_j(t). \tag{4-5}$$

4.2.2 Proposed classification model with an unlearned-class probability density function

Anomalous time-series data not assumed in the training phase can be interpreted as not following the probability density function determined from training for all states. The incorporation of PDFs for unlearned classes into the state of the HSMM allows estimation of the probability of not belonging to trained states, and newly included states can be labeled as unlearned. In this approach, the complementary Gaussian distribution (CGD) technique developed by Shima *et al.* [21] was employed. Based on the assumption that input data follow Shima's CGD pattern, data in the unlearned class are distributed around each component. From these definitions, the output probability distribution of the unlearned state is defined from the linear sum of CGD as

$$b_{K+1}(\boldsymbol{x}_t; \bar{\boldsymbol{\phi}}_k) = \frac{1}{\sum_{k'=1}^{K} M_{k'}} \sum_{k=1}^{K} \sum_{m=1}^{M_k} h\left(\boldsymbol{x}_t; \bar{\boldsymbol{\phi}}_{k,m}\right). \tag{4-6}$$

Here, $K + 1$ is the number of unlearned states, and the parameter set $\bar{\boldsymbol{\phi}}_k = \{\bar{\boldsymbol{\phi}}_{k,m}\}_{m=1,\cdots,M_k}$ is defined. Optimization of the new set of HSMM parameters in consideration of the unlearned state $\boldsymbol{\theta}_{\mathrm{new}} = \{\pi'_k, \boldsymbol{A}', P'_k(d), \bar{\boldsymbol{\phi}}_k\}_{k=1,\ldots,K+1}$ is needed to enable hidden-state estimation and anomaly detection using (4-5).
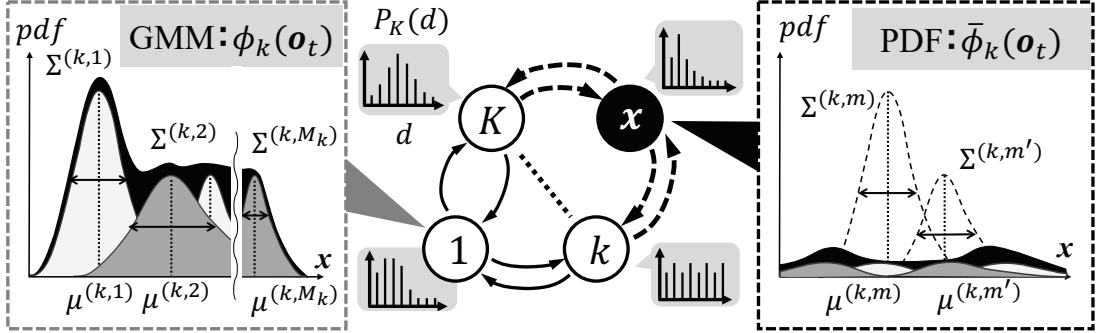
Fig. 4.1: Hidden semi-Markov model with the unlearned state

4.2.3 Model parameter estimation for the proposed hidden semi-Markov model

In the HSMM training process, the parameters of $\hat{\boldsymbol{\theta}}$ are optimized to maximize likelihoods $\hat{\boldsymbol{\theta}}$ for measurement signals $\boldsymbol{X}$. In the work reported here, the EM algorithm is used to train normal states other than the unlearned state. The forward-backward algorithm proposed by Yu $et\ al.$ [23,24] was employed to avoid very low values with underflow resulting from likelihood computation using long time-series data.

For the HSMM pre-trained using the EM algorithm, an unlearned state is additionally applied with unoptimized parameters. However, as training data sets for unlearned classes are not provided, $\hat{\boldsymbol{\theta}}_{\mathrm{new}}$ cannot be estimated using the Baum-Welch algorithm. Thus, it is necessary to estimate parameters for the unlearned state using the training data $\boldsymbol{X}$, which do not include anomalies. First, it is assumed that the transition of probability to an unlearned state in the initial state is very small and is defined as $\pi'_{K+1} = p_{\pi'}$, where $p_{\pi'}$ is an actual constant close to 0. Then, assuming that outliers do not occur frequently, the probability of transition to the unlearned state $a'_{i,K+1}$ contained within the matrix of state transition probabilities $\boldsymbol{A}' \in \Re^{(K+1)\times(K+1)}$ can be determined as

$$a'_{i,K+1} = a_{\mathrm{const}} \tag{4–7}$$

$$a'_{i,j} = (1 - a_{\mathrm{const}})a_{i,j} \quad (i, j = 1, \cdots, K). \tag{4–8}$$

Here, $a_{\mathrm{const}} < 1$ is a constant. As there is no prior information relevant to transitions from the unlearned state to normal states, the probability of transition from the unlearned state $a'_{K+1,j}$ is calculated from

$$a'_{K+1,j} = \frac{a_j^*}{\sum_{i=1}^{K} a_i^*} \tag{4-9}$$

$$a_j^* = \sum_{i=1}^{K} a_{i,j} \ (j = 1, \cdots, K) \tag{4-10}$$

$$a'_{K+1,K+1} = 0. \tag{4-11}$$

In addition, assuming that the time remaining in the unlearned state is brief, $P'_{K+1}(\cdot)$ is modeled via gamma distribution.

Finally, the parameters of the PDF h of unlearned classes $\bar{\phi}_{k,m}$ is optimized. As $\bar{\phi}_{k,m}$ is composed of $\phi_{k,m}$ and $\epsilon_{k,m}$, h based on the parameters of $\phi_{k,m}$ can be used to represent anomaly distribution. Thus, $\bar{\phi}_{k,m}$ can be determined via the optimization of $\epsilon_{k,m}$ using training samples. In the training of $\epsilon_{k,m}$, $p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}_{\mathrm{new}})$ is calculated for normal training data, and hidden-state estimation with the unlearned state can be performed as

$$\hat{s}'(t) = \arg \max_{j \in K+1} \alpha'_j(t). \tag{4-12}$$

As the training data do not include transition to an unlearned state, $\epsilon_{k,m}$, which allows a reduction in the number of training data recognized for the state $N_{\hat{s}=K+1}$, is required. Thus, using gradient descent, $\epsilon_{k,m}$ is updated until $N_{\hat{s}'=K+1}/T \leq N_{\mathrm{th}}$ is satisfied as

$$\epsilon_{q,k,m}^{\mathrm{new}} = \epsilon_{q,k,m}^{\mathrm{old}} + \Delta_q \ (\Delta_q > 0). \tag{4-13}$$

Here, N_{th} is an actual constant close to 0, and Δ_q $(q = 1, 2)$ represents update coefficients.

4.2.4 A multi-class classifier based on the proposed hidden semi-Markov models

The proposed HSMMs incorporating unlearned states are trained for all classification target events, and pattern recognition is achieved based on Bayesian discrimination. The *a posteriori* probability of class c, $p(c|\boldsymbol{X})$ ($c = 1, \cdots, C$), can be calculated from the likelihood of each class $p(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{new}}^c)$ using

$$p(c|\boldsymbol{X}) = \frac{p(c)p(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{new}}^c)}{\sum_{c'=1}^{C} p(c')p(\boldsymbol{X}|\boldsymbol{\theta}_{\mathrm{new}}^{c'})}. \tag{4–14}$$

Here, C is the number of classes. Given that each event is dependent, multiclass classification is enabled by selecting the class with the maximum *a posteriori* probability based on Bayesian discrimination. After the assignment of class labels to the input time-series data, anomaly detection with the unlearned state is performed on the data for each time.

For anomaly detection, $\alpha'_{c,j}(t)$ ($c = 1, \cdots, C$, $k = 1, \cdots, K + 1$) as determined via likelihood calculation for each class is utilized. As per Eq. (4–12), the transition sequence of the hidden state $\hat{s}'_c(t)$ can be estimated from $\alpha'_{c,j}(t)$. If $\hat{s}'_c(t) = K + 1$ is satisfied, $\boldsymbol{x}_t$ is classified as an unlearned state not belonging to class c. If transition to the learned normal state is labeled as 0 and transition to the unlearned state is labeled as 1, a binary signal $f_c(t)$ can be obtained using

$$f_c(t) = \begin{cases} 0 & (\hat{s}'_c(t) \neq K + 1) \\ 1 & (\hat{s}'_c(t) = K + 1). \end{cases} \tag{4–15}$$

The conjunction is performed at each time between classes, and the results of anomaly detection $F(t)$ are determined as

$$F(t) = \prod_{c'=1}^{C} f_{c'}(t). \tag{4–16}$$

If $F(t) = 1$ is satisfied, $\boldsymbol{x}_t$ belongs to the unlearned class and is classified as an outlier. Combining the results of multiclass classification and hidden-state sequence estimation enables accurate multi-class discrimination and detailed anomaly detection.

4.3 Experiments

4.3.1 Classification experiments for artificial signals

To evaluate the performance of the proposed method, artificial data classification experiments were conducted for two-dimensional artificial signals incorporating four learned classes and an unlearned class. To generate training time-series data, ED-HMMs with GMMs were employed. Predefined ED-HMMs can be used to produce complex signals based on the semi-Markov chain and output PDFs. Each ED-HMM has four states, each with a GMM consisting of four Gaussian distributions, and the maximum duration value was set as 500.

In test data generation, new ED-HMMs with unlearned states added to the ED-HMM applied for training data generation were used to evaluate anomaly detection performance. The new ED-HMM set-up has four normal and four abnormal states incorporating GMMs. As the added unlearned states are common to all classes, the test data generated are signals with four abnormal values added to normal time-series data (Fig. 4.2). Class transitions occur randomly in line with predefined transition probabilities. Here, the probability of outliers was about 20%, and uniform random numbers were used to generate the model parameters.

For the learned classes, a total of 8,000 samples (2,000 for each class) were used as training data, and the model parameters for the training were $K = 3$, $M_k = 2$, $D = 200$. The duration density of the unlearned state was modeled on gamma distribution with a value of 10 for both the shape parameter and the scale parameter, and $\epsilon_{k,m}$ was updated until N_{th} fell below 0.01. For classification performance evaluation, 10,000 evaluation data values containing anomalies were generated and classified using the proposed method based on Eqs. (4–14) and (4–16). The length of the signal input to the classifier was set to 100 samples, and the overlap with the signal to be input next was set to 90 samples.

To validate the capacity of the approach in producing accurate classification, the results were compared with those from previous methods. For anomaly de-

tection with time-series data, the method previously proposed by the authors [37] was employed because it has similar training conditions appropriate for performance comparison. A multi-class SVM along with a single-class SVM were also used for anomaly detection in addition to the ED-HMM with the GMM [23, 24]. The single-class SVM allows state-of-the-art anomaly detection, and is often used as a comparative approach to semi-supervised training problems normally associated with non-anomalous data [33]. The SMV-based classifier employed a radial basis function (RBF) kernel, and the model structure of the comparative method was as per the proposed technique ($K = 3, M_k = 2$). To validate the capacity of the approach in producing accurate classification, the results were compared with those from the previously proposed method.

Figure 4.2 details classification results produced using the proposed and previous methods (from top to bottom: true class label, generated evaluation signals, classification results obtained with the proposed method and those obtained with the previous method). C0 indicates data classification to the unlearned class as outliers, and grey areas represent abnormal parts in test data. The outcomes indicate that both learned and unlearned classes can be classified using the proposed method. However, although both are essentially recognized using the previous method, many misidentifications in unlearned classes occur after inter-class transition. This may be partially attributable to the utilization of classification for the unlearned class based on *a posteriori* probability. If the input signal contains boundaries between classes or instantaneous outliers, the reduced likelihood of each class will cause ambiguous discrimination results. As the previous method involves estimation regarding the degree of abnormality for entire input signals, time-series data with an ambiguous class will belong to the unlearned category. In contrast, as the proposed method involves the performance of anomaly detection at each time, even short anomalies can be classified with high accuracy.

Figure 4.3 shows average classification ratios for all classes and standard errors. The proposed method produced high ratios of 95.7% for evaluation data and 99.1% for the unlearned class. The corresponding values were 88.8% and

92.0% with the previous method and 87.31% and 92.9% with the SVM classifier. The outcomes showed that a multi-class classifier based on ED-HMM achieved the highest classification ratio for learned classes, but was unable to identify motions of unlearned classes. The results demonstrated that the proposed method enabled anomaly detection with no degradation of discrimination performance for learning classes as compared to an ED-HMMs classifier. In addition, a significant difference between the proposed method and the SVM was observed. The recognition performance of the comparison methods exhibited some degeneration, with the previous method not allowing consideration of state duration and the SVM not allowing the use of information from time-series data.

4.3.2　Human action recognition for simulated care tasks

To demonstrate the effectiveness of the proposed approach with real data, motion classification for simulated care tasks was conducted. Due to Japan's rapid demographic aging, the heavy burden placed on care workers has become a major issue in the country. Accordingly, there is a need to reduce the workload of such staff and analyze the roles involved in personal care work.

Individual care involves a variety of small operations as shown in Table 4.1, which details a chair-to-bed transfer sequence. As the major operations here (such as assistance for a standing posture) are predefined among related care tasks, motion classification can be achieved via classifier training. However, the definition of connections between individual actions is ambiguous, and often includes unpredictable actions such as unexpected movement and squatting. If conventional methods are used for observation in actual care work, unclassified motions may adversely affect recognition performance. In the part of the study reported here, minor unexpected movements were recognized to verify the capacity of the proposed method.

In the experiment, care task data based on simulation of transfer from a wheelchair to a bed were recorded via motion capture, and classification for minor operations was performed. Figures 4.4 and 4.5 show the experimental en-
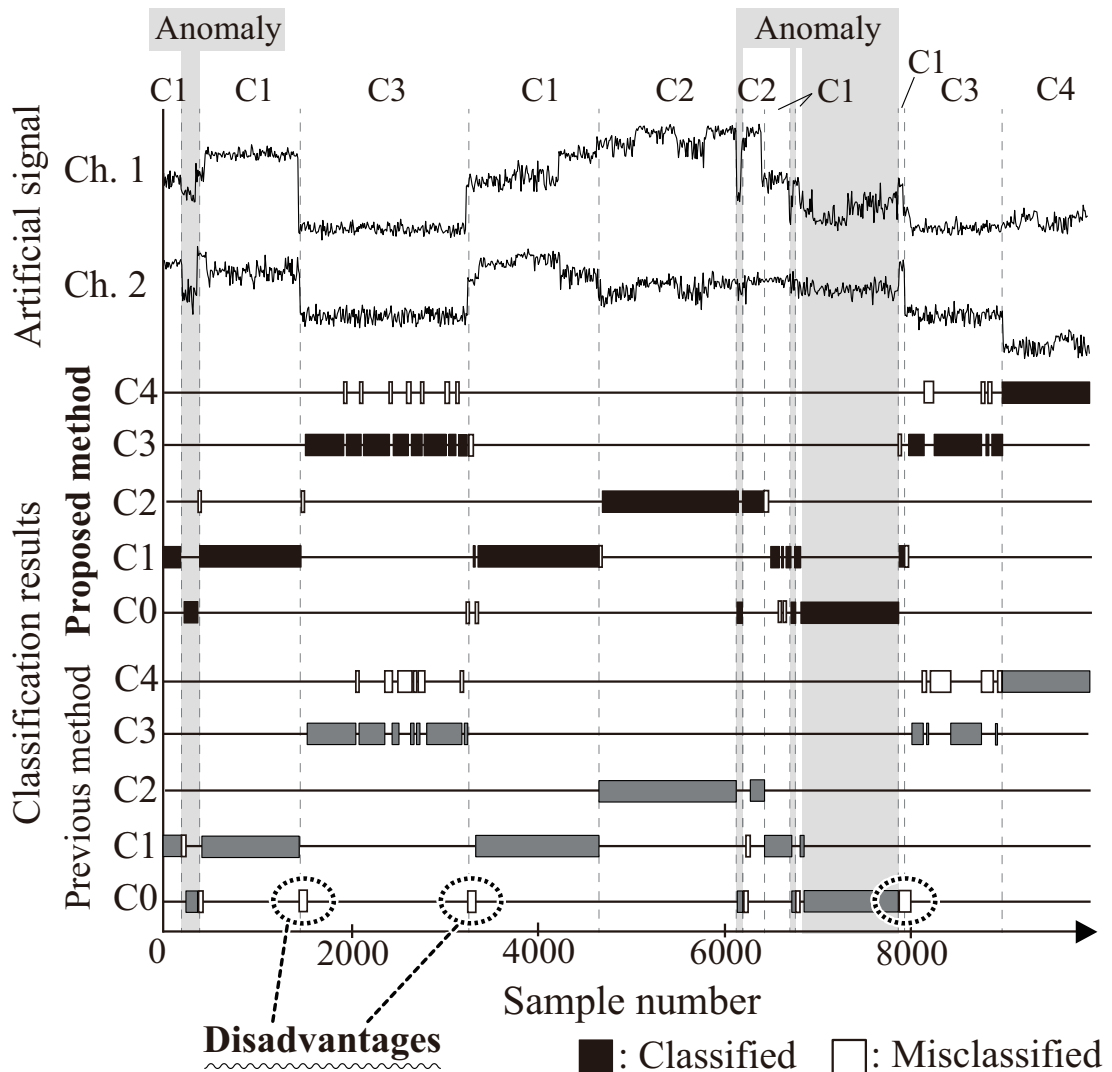
Fig. 4.2: Artificial data generated from ED-HMM and classification results

vironment and an example of the simulated care operation, respectively. In the simulated tasks, a trolley was used as a wheelchair, and a water bag (WB) was treated as the care recipient. The tasks were implemented in a fixed order with the objective of laying the WB on a table (Table 4.1). Data for three trials were recorded from one subject, and the posture angles of each body link determined from motion capture were used for training and evaluation. Principal component analysis (PCA) was employed to extract characteristics from 54-dimension data,
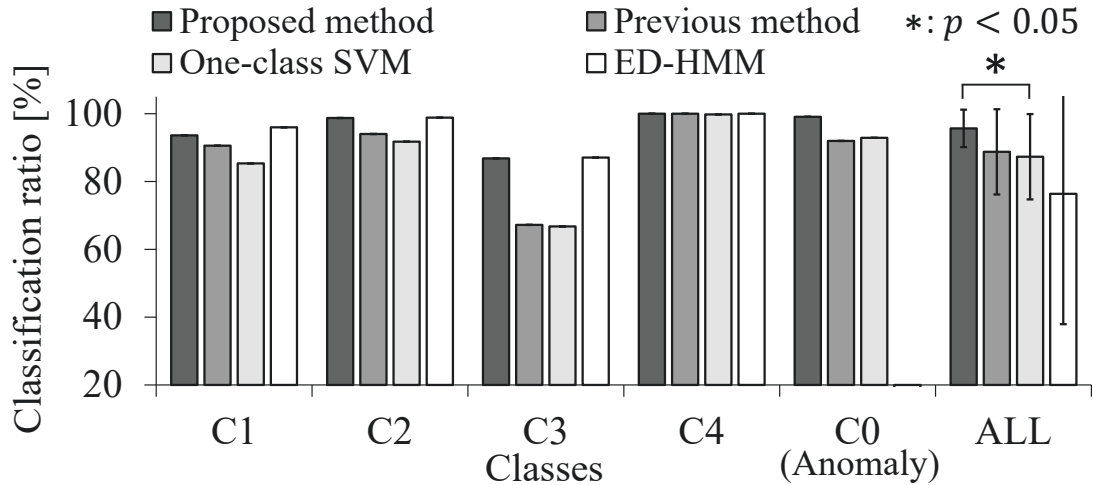
Fig. 4.3: Classification ratios of anomaly detection methods

and 7-dimension characteristics were generated to enable ED-HMM training with a contribution ratio of approximately 80%. The proposed method for each minor operation was applied for training based on the definitions and recorded labels for simulated care tasks, with model parameters set as $C = 5$, $K = 2$, $M_k = 1$ and $D = 100$. Although the training data included unclassified motions, training for the proposed model utilized only normal time-series data with unexpected motions excluded. Threefold cross-validation for two trials and test data for one trial were implemented to evaluate classification with the proposed method as compared with the previous method and the single-class SVM with RBF kernel application.

Figure 4.6 shows average classification ratios (based on six classes, including the unlearned one, represented by "Ave.") and standard deviations determined from cross-validation. The results indicate that the proposed method produced better classification ratios for cross-validation and exhibited stabler performance. In addition, significant differences were observed between the proposed method and the comparison method. However, the standard deviation in each trial appears large, and incorrect learning was seen with some tasks. Figure 4.7 shows the average classification ratio for each operation and the related standard deviation

Table 4.1: Wheelchair-to-bed transfer

| Practical care tasks	Simulated care tasks	Motion label
Push a wheelchair	Push a dolly cart	E1
Put a brake	Move	
Undefined action （Move etc.)	Crouch down	E0
Foot support operation	Move weights	E2
Undefined action	Stand up & move	E0
Sitting assistance	Transfer WB into the frame	E3
Assistance for standing up	Change the handle of BW	E5
Move the wheelchair		
Assistance to lie in a bed	Lay WB on the table	E4
Raise fences of the bed	Move	E0
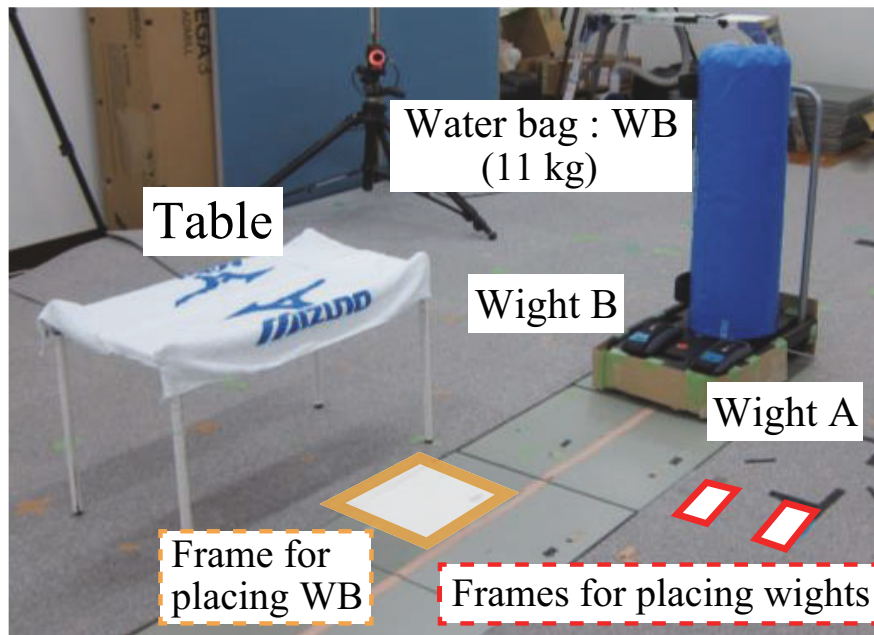
Fig. 4.4: Experimental environment

obtained from cross-validation. The performance of the proposed method is deteriorated in E4 and E0, and significant variance was observed. This indicates that classification performance largely depends on the training data used, and that a lack of appropriate training data can adversely affect the discrimination ratio. However, it is also seen that the unlearned class with the previous method is not

(a) Push a dolly cart: E1   (b) Move weights: E2

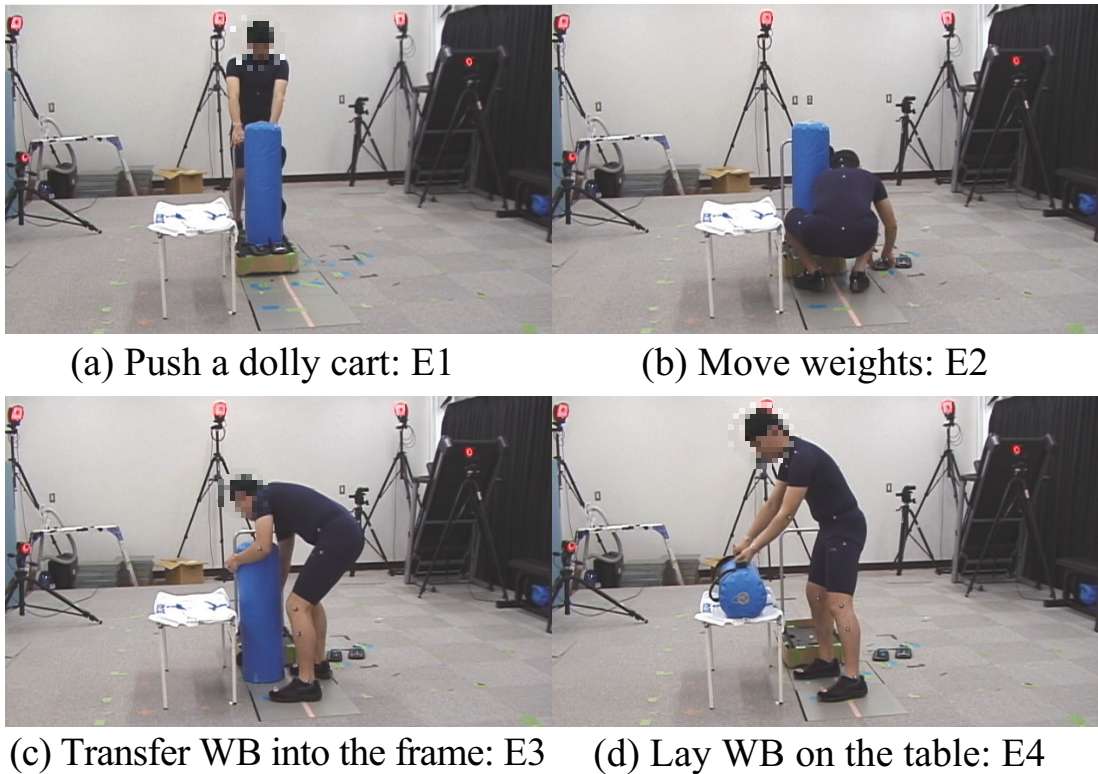(c) Transfer WB into the frame: E3   (d) Lay WB on the table: E4

Fig. 4.5: Examples of simulated transfer operations

classified at all, and the classification ratio of learned classes was significantly deteriorated with the SVM method. These results indicate the effectiveness of the proposed method for data whose handling was problematic with the previous method. In future work, the authors plan implementation with a larger body of learning data and improved identification accuracy.

4.3.3 Evaluation of work recognition performance for actual care tasks

Posture monitoring and motion detection experiments were conducted in an actual work environment to evaluate the classification performance of the proposed method in discrimination involving actual care tasks.

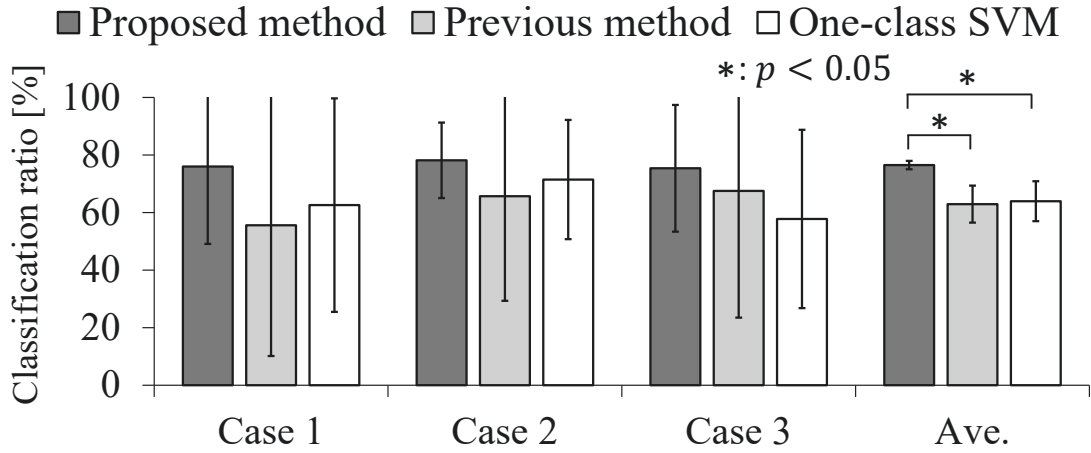As methods involving multiple cameras (such as motion capture systems and

Fig. 4.6: Classification ratios from threefold cross-validation
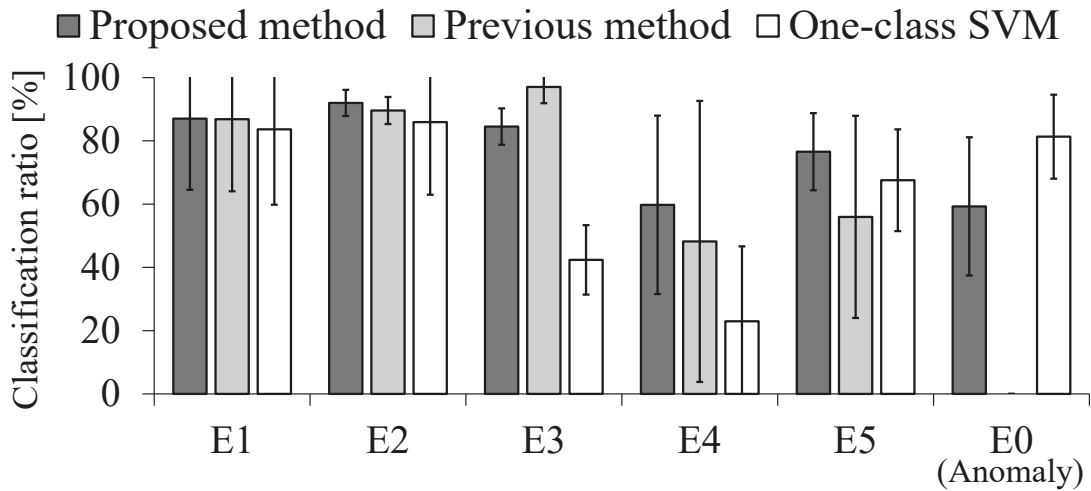
Fig. 4.7: Classification ratios for individual motions

Kinect) that are widely used in human behavior monitoring [49–51] are inappropriate for use in nursing homes due to issues with privacy, obstacle avoidance and difficulty in tracking movement, the proposed system involves posture monitoring using wearable inertial sensors (Figure 4.8 (a)). These determine acceleration, angular velocity and orientation, providing data that can be used to calculate inclination in three-dimensional space (i.e., the spatial relationships of monitoring points). The sensors are small and light, enabling attachment within clothes for

low-load monitoring.

As the installation of inertial sensors all over the body for detailed monitoring may interfere with caregiver tasks, the number of monitoring points should be minimized for accuracy of work classification. In preliminary experiments, detailed monitoring with N inertial sensors across the body was conducted. N_{ini} links with high-posture-angle monitoring accuracy were selected from the resulting data, and a work discriminator for target tasks was trained from the characteristics of the selected regions. Discriminative performance was then evaluated using the test data, and N_{ini} links for target values with high classification accuracy were determined as sensor attachment positions. Where the target accuracy was not reached, an unselected region was chosen and learning accuracy was verified via addition to the relevant characteristics. One link among the unselected options was also added to the characteristics, and learning accuracy was verified again. This re-validation was performed for all unselected links, and the link with the highest classification accuracy was added to the monitoring points. The process was repeated as necessary to determine appropriate sensor location.

In the work classification process, a new hierarchical HSMM was applied to the wheelchair-to-bed transfer task, which is a major operation involving a large physical load. The HSMM consists of two layers and is suitable for classification of small to large operations. The first layer models individual minor care tasks with an unlearned state, and the second calculates the probability that a transfer operation has been performed from the small-operation sequence determined in the first layer. The number of states in the second layer matches that of target small operations, and categorical distribution is introduced as output.

A monitoring experiment on care-giver work postures was conducted with two subjects in a nursing home (Ashihara Nozomien, Fukuoka), and wheelchair-to-bed transfer was evaluated offline. Inertial sensors were fitted at six locations determined from preliminary experiments during monitoring (Figure 4.8 (b)). Posture information obtained from the sensors was expressed in 24 dimensions, and 16-dimensional characteristics were determined via PCA. The task labels used

<div align="center">(a) Used inatial sensor   (b) Location of sensors</div>
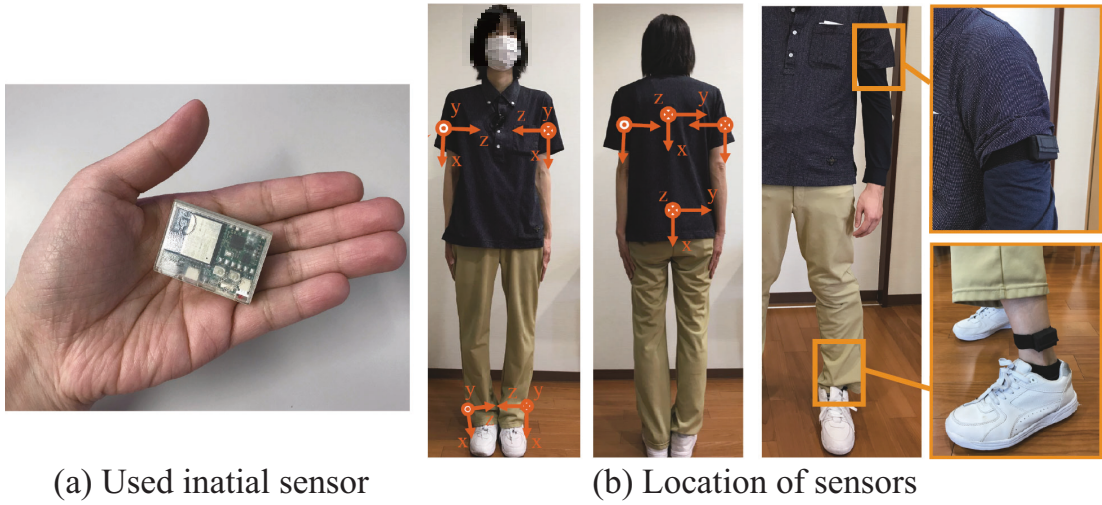
Fig. 4.8: Posture monitoring with wearable inertial sensors

in training were based on observation of actual nursing care tasks by experts in the field. Table 4.2 shows classification targets in minor operation relating to wheelchair-to-bed transfer, and other operations were identified as unlearned. The HSMM parameters in the first layer were set as $K = 3, M_k = 2$ and $D = 200$, and the number of states in the second layer was $K' = 6$. The training data for the first layer excluded unlearned tasks, and the test data contained all monitoring information, including unlearned/unclassified tasks. Small-operation classification was implemented, and the probability of transfer assistance was calculated every 0.1 seconds.

Figure 4.9 (a), (b) shows transfer task results for Subjects A and B, respectively (from the top: input characteristics (three dimensions) and transfer task probability determined from the HSMM of the second layer). Shaded areas represent periods during which transfer assistance was provided. The results for Subject A indicate that all transfer operations were correctly identified, while those for Subject B contained many misidentifications and probabilistic ambiguities. These outcomes help to indicate the effectiveness of the proposed method. The lower classification capacity observed may be attributable to the work rate of Subject B, who was observed to perform nursing care tasks significantly faster

Table 4.2: Bed-to-wheelchair transfer

(a) Routine of practical care tasks

Push a wheelchair
Put a brake
Undefined action (Move etc.)
Foot support operation
Undefined action
Sitting assistance
Assistance for standing up
Move the wheelchair
Assistance to lie in a bed
Raise fences of the bed

(b) Learned and Unlearned motions

Learned	①	Shoes & Foot support operation
	②	Assistance for standing up
	③	Assistance to lie in a bed or get up from bed
	④	Postural change
	⑤	Sitting assistance
	⑥	Jack operation
Unlearned		Move the wheelchair
		Operation for the fence of bed
		Walk
		Other…

than Subject A. Accordingly, the transition between small tasks was unclear, making it difficult to record task labels. As a result, labels different from the actual work were created in the training data, causing reduced discrimination accuracy for both layers. It can also be inferred that monitoring performance was affected by the small training data set, variations in the degree of care required by the recipient, and inappropriate manually set hyperparameters.

In future research, it will be necessary to develop new learning regulations based on log-linearization for appropriate training from a small training data set, to create a general classification model that does not require individual training, and to consider increasing the number of detectable actions.

4.4   Concluding remarks

Section 4 outlines a novel HSMM-based sequential pattern recognition method incorporating unlearned states to represent unexpected anomalies in the learning process based on time-series data characteristics. The approach provides high classification performance in consideration of time-series information relating to signals, and allows more detailed anomaly detection than the previous method. In the study, two pattern recognition experiments were performed to evaluate
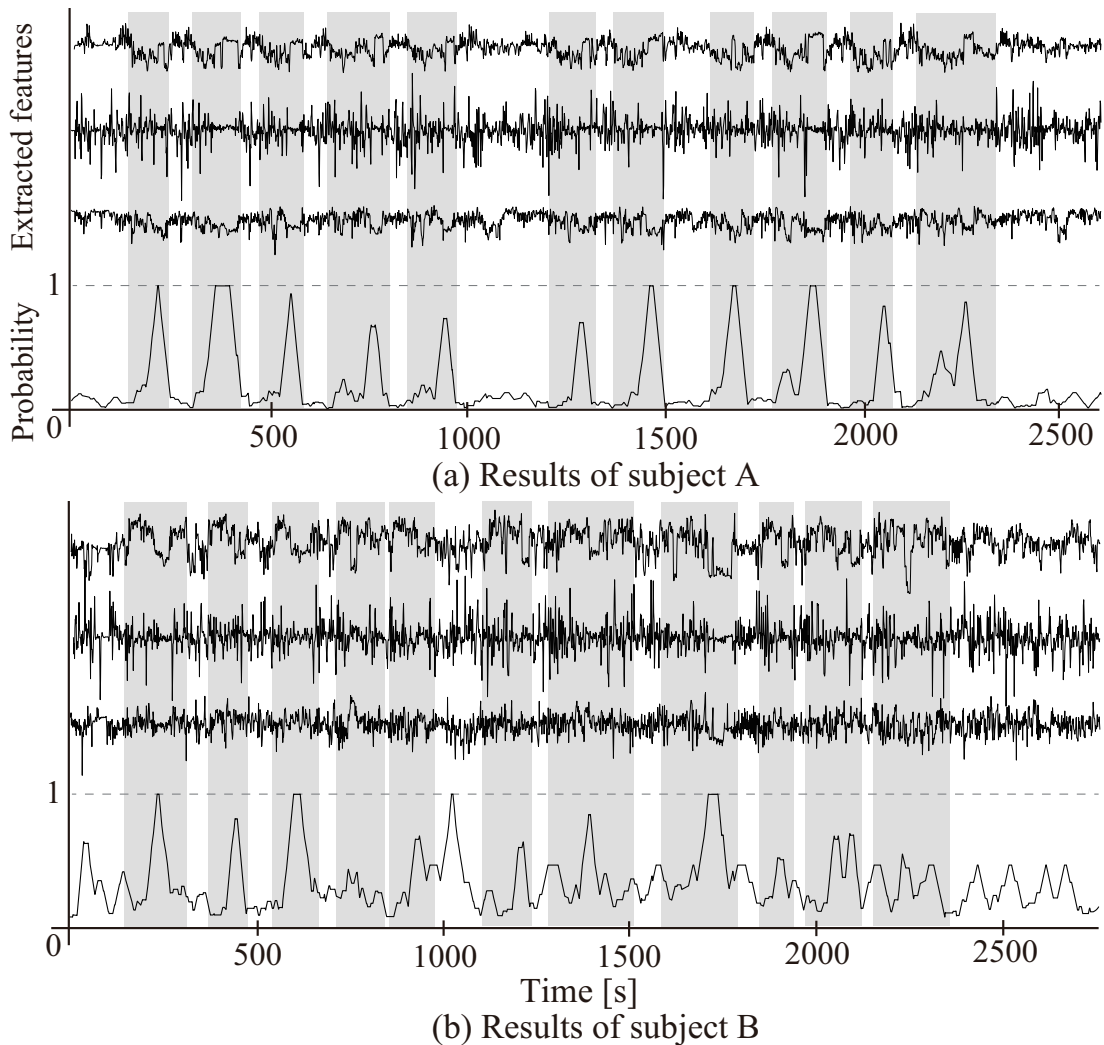
Fig. 4.9: Probability calculation results for transfer assistance work

the technique. In artificial signal classification, generation of complex time-series data from ED-HMMs indicated the method's suitability for this purpose. In experiments relating to simulated care tasks, motion classification performed with characteristics determined from motion capture demonstrated the method's superiority over other approaches.

To further highlight the competitive performance of this technique, additional evaluation will be performed with artificial signals generated from various time-series models, a greater number of subjects and application to other anomaly

detection problems.

Chapter 5

Conclusion

This paper proposes novel unlearned class determination based on complementary Gaussian distribution to overcome the limitations of conventional anomaly detection. To improve training performance with small datasets, a novel probabilistic neural network (NACGMN) for multi-class discrimination incorporating unlearned classes with a single classifier was developed. The proposed approach involves learning with relaxed statistical constraints based on a back-propagation algorithm and stochastic parameter extraction with improved network readability. To expand the range of application for the network, unlearned pattern recognition based on a hidden Markov model (OVRHMM) was proposed for time-series data classification. To enhance anomaly detection performance for time-series data, a hidden semi-Markov model incorporating an unlearned state and novel pattern classification involving the proposed HSMMs were also developed. These approaches were applied to classification problems with specialized artificial data and biological signals to validate performance.

Section 2 outlines the probabilistic neural Normal and Complementary Gaussian Mixture Network (NACGMN) for determination of *a posterior* probability for unlearned classes. The network incorporates the GMM and CGMM, which represent the distribution of training samples for each class and the imaginary distribution of unlearned classes, respectively. In the NACGMN structure, the

stochastic parameters of the GMM and CGMM are converted to unconstrained weight coefficients based on log linearization [5], and can be optimized using the back-propagation algorithm and applied to semi-supervised learning only with normal training data free of abnormal samples. Stable learning can thus be achieved even with small data sets. Parameter extraction for the NACGMN was also developed, and efforts were made to improve readability. The superiority of the proposed method was verified from two experiments. In artificial data classification, the NACGMN was found to support superior classification even with a small number of training samples, and enabled more stable decision than comparative methods. It was also shown that the GMM and CGMM approach enabled extraction from a trained NACGMN with a similar decision region. In motion classification experiments, the approach was applied to recognition of EMG signals from seven subjects each performing eight forearm motions. NACGMN classification ratio accuracy was significantly higher than those of other methods for all subjects. These outcomes demonstrated the effectiveness of the proposed technique.

Section 3 outlines novel hidden Markov model-based sequential pattern recognition (OVRHMM) allowing identification of unexpected unlearned classes in the learning process based on time-series data characteristics. The OVRHMM is an extension of the unlearned class detection proposed by Shima [21] to time-series data classification, and enables evaluation of the degree of anomaly for all input signals. In this model, Bayesian discrimination for unlearned classes is realized based on likelihood calculated from HMMs and complementary HMMs. To validate the performance of the approach, three pattern recognition experiments were conducted. In artificial signal classification, the time-series information-based OVRHMM was found to enable superior classification in anomaly identification. The outcomes of motion classification experiments also demonstrated that the proposed approach allows classification of unlearned classes with no deterioration in skill for learned classes, and that its performance was superior to those of other methods except the NACGMN. The approach was also applied to anomaly

identification for neurodegenerative conditions using gait data. In the scenario of this experiment, subjects with neurodegenerative conditions were placed in the unlearned class for training without disorder gait data. The technique was found to demonstrate a level of performance similar to that of the previous method, in which disorder data are used for training. These results indicated the effectiveness of the complementary event model for time-series data and the expandability of the NACGMN.

Section 4 discusses the validation of HSMM-based sequential pattern recognition incorporating unlearned states to represent unexpected anomalies in the learning process based on time-series data characteristics, with the objective of OVRHMM improvement. The approach cannot be used to identify abnormalities in time-series data, and the clarity of identification results may depend on the ratio of abnormal values included. The proposed approach helps to resolve these issues and provides high classification performance in consideration of time-series information relating to signals, as well as allowing more detailed anomaly detection than the previous method. In the artificial signal experiment, the approach produced less misidentification than the OVRHMM, indicating its suitability for this purpose. In experiments with simulated care tasks, motion classification performed with characteristics determined from motion capture demonstrated the method's superiority over other approaches. The results indicated that the approach involves the use of a sufficiently general-purpose stochastic model structure for anomaly detection in time-series data. It should be noted that the method incorporates a number of points that will require improvement in future work.

Although this paper discusses the probabilistic anomaly detection method based on complementary event models and the training method, it should be noted that the proposed approach incorporates a number of points that will require improvement in future work.

As reported in Section 2, learning was significantly slower than that of the EM algorithm because the proposed NACGMN employs the constrained error back-propagation method. High-speed parameter optimization is required for

implementation with myoelectric prosthetic hands, either with improved parameter updating or faster hardware. Adaptive adjustment of the learning rate is an effective way to speed up parameter updates, and this can be achieved with the AdaGrad algorithm [52] and a terminal attractor [53]. Research on GMM hardware acceleration [54, 55] has also indicated the effectiveness of implementation on the field-programmable gate array (FPGA). In future work, it will be necessary to consider faster learning from both software and hardware based on these methods.

In consideration of discussions regarding issues with the proposed method, the authors plan to revalidate the characteristics of the NACGMN for high discrimination capacity with very limited training data sets. In application to biological signal classification with few training samples, work is needed to examine the network's applicability to classification in regard to individual differences. Further work is also needed to clarify the versatility of the NACGMN with learning from individuals and the trade-off between the number of samples in the target cluster and NACGMN capacity. One approach in this regard may involve reusing a NACGMN model trained for a particular subject for other subjects in motion classification using EMG signals.

Sections 3 and 4 highlight the difficulty of developing training with relaxed statistical restrictions for the HMM and HSMM. Stable estimation of covariance matrix content for real data is extremely challenging, but this can be addressed with the addition of a correction term during parameter estimation. However, as this process adversely affects identification accuracy, learning based on log-linearization is required. Since the outcomes described in Section 4 demonstrate the effectiveness of HSMMs with unlearned states, a new stochastic neural network needs to be developed based on this structure. For more efficient learning, the advance determination of appropriate hyperparameters (such as the number of states and components) is important. Future work will also involve consideration using an infinite GMM [56] and an infinite HMM [57] for pre-learning.

As the proposed approach enables only determination of the presence or ab-

sence of abnormalities in the application of technology for disease diagnosis, the structure of the unlearned class involves a simple black-box technique, and the characteristics of outliers cannot be known. To address this, analysis of structure from samples in the unlearned class is required. As no prior information on unlearned classes can be obtained, future work will involve application of the nonparametric Bayes method for learning hyperparameters [56, 57].

The proposed method is regarded as an anomaly detection approach, and is intended for application to diagnostic support for biological signals and medical data as well as machine failure detection. In classification with innumerable discrimination targets using this approach, only certain important elements are identified and other events are placed in the unlearned class, thereby avoiding significantly reduced discrimination accuracy. Related issues include human motion classification, for which the experiment results outlined here show the method's effectiveness. In future work, the approach will be applied to automatic generation of action records in combination with wearable sensors and stable control of myoelectric prosthetic hands.

To improve classification performance with the proposed method, future work will also include the development of a classification model integrating the NACGMN and OVRHMM unlearned class detection approaches, with the integrated classifier defined as a novel probabilistic recurrent neural network capable of considering unlearned class data. In contrast to regular novel technique development, the introduction of expertise on conventional neural networks to the NACGMN is also important. As the NACGMN model is more applicable to the handling of low-dimensional data, it requires reduction for high-dimensional input. Hence, the introduction of a convolution layer to enable feature extraction, dimension reduction, handling of image data and a wider application range is also desirable. The capacity for end-to-end learning from the feature extraction section to the *a posteriori* probability calculation section via model integration will also enable fine-tuning of the encoder section during NACGMN training. Accordingly, the generation of feature vectors that are difficult to misidentify in the

unlearned class will be possible. Consideration of various functional extensions based on additional PDF introduction is expected to support the proposed unlearned class estimation for use as a classifier providing superior versatility and discrimination performance. Gamma distribution, Rayleigh distribution, Johnson SU distribution and others are candidates for expansion, but the definition of complementary event distribution for each PDF is a significant issue, and the linearization of complex functions such as the gamma function is important. Future work will require the development of a new neural network technology that supports simple understanding via probabilistic-model introduction and fusion.

A Appendix: Complementary Gaussian distribution

In the proposed method, the distribution of abnormal samples unexpected in the training phase was defined as

$$
h\left(\boldsymbol{x}; k, m\right) = D^{-1}(2\pi)^{-\frac{D}{2}} \varepsilon_{k,m}^{-1} \left|\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}\right|^{-\frac{1}{2}}
$$
$$
\times \left(\boldsymbol{x} - \boldsymbol{\mu}_{k,m}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{k,m}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_{k,m}\right)
$$
$$
\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_{k,m}\right)^{\mathrm{T}}(\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m})^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_{k,m}\right)\right\}. \tag{A-1}
$$

By substituting $q(\boldsymbol{x})$ as defined in Eq. (2–3) into this definition, Eq. (2–6) can be derived.

Here, the antiderivative of $h\left(\boldsymbol{x}; k, m\right)$ can be derived for $\boldsymbol{x}$. When $\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{\mu}_{k,m}$ and $\boldsymbol{S} = \varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}$ are set, the antiderivative is expressed as

$$
\int h\left(\boldsymbol{x}; k, m\right) d\boldsymbol{x}
$$
$$
= A\varepsilon_{k,m} \int d\boldsymbol{x}' \, \boldsymbol{x}'^{\mathrm{T}} \boldsymbol{S}^{-1} \boldsymbol{x}' \exp\left(-\frac{1}{2}\boldsymbol{x}'^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{x}'\right). \tag{A-2}
$$

Here, A is defined as

$$
A = D^{-1}(2\pi)^{-\frac{D}{2}} \varepsilon_{k,m}^{-1} \left|\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}\right|^{-\frac{1}{2}}. \tag{A-3}
$$

As $\boldsymbol{S}$ is symmetric and positive semi-definite, the eigenvalues of $\boldsymbol{S}$: $\{\lambda_1, \cdots, \lambda_D\}$ satisfy $\lambda_d \geq 0$ $(d = 1, \cdots, D)$. Using an orthogonal matrix $\boldsymbol{\Phi} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_D\}$, $\boldsymbol{S}^{-1}$ can also be diagonalized as

$$
\boldsymbol{R} = \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{S}^{-1} \boldsymbol{\Phi}
$$
$$
= \mathrm{diag}\left(\lambda_1^{-1}, \cdots, \lambda_D^{-1}\right) \tag{A-4}
$$

where $\boldsymbol{v}_d$ $(d = 1, \cdots, D)$ represents eigenvectors of $\boldsymbol{S}^{-1}$. When $\boldsymbol{z} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}'$ is set,

Eq. (A–2) can be transformed as

$$A\varepsilon_{k,m} \int d\boldsymbol{z} \ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{z} \exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{z}\right)$$

$$= A\varepsilon_{k,m} \int d\boldsymbol{z} \ (\lambda_1^{-1}z_1^2 + \cdots + \lambda_D^{-1}z_D^2) \exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{z}\right)$$

$$= A\varepsilon_{k,m} \left\{ \sum_{d=1}^{D} \int d\boldsymbol{z} \ \lambda_d^{-1}z_d^2 \exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{z}\right) \right\}. \tag{A–5}$$

Consideration of the dth element produces

$$\int d\boldsymbol{z} \ \lambda_d^{-1}z_d^2 \exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{z}\right)$$

$$= \lambda_d^{-1} \int_{-\infty}^{\infty} z_d^2 \exp\left(-\frac{1}{2}\lambda_d^{-1}z_d^2\right) dz_d \times \prod_{i\neq d} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\lambda_i^{-1}z_i^2\right) dz_i$$

$$= \lambda_d^{-1} \left(\lambda_d\sqrt{2\pi\lambda_d}\right) \prod_{i\neq d} \sqrt{2\pi\lambda_i}$$

$$= (2\pi)^{\frac{D}{2}} |\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}|^{\frac{1}{2}}. \tag{A–6}$$

These results can be substituted into Eq. (A–5) to prove

$$A\varepsilon_{k,m}D(2\pi)^{\frac{D}{2}} |\varepsilon_{k,m}\boldsymbol{\Sigma}_{k,m}|^{\frac{1}{2}} = AA^{-1}$$

$$= 1. \tag{A–7}$$

Thus, $\int h(\boldsymbol{x}; k, m) \, d\boldsymbol{x} = 1$ is met. When $\varepsilon_{k,m} > 0$ is met, $A > 0$, $-2q(\boldsymbol{x}) \geq 0$ and $\exp[q(\boldsymbol{x})] > 0$ are also met. Hence, $h(\boldsymbol{x}; k, m) \geq 0$ can be met for arbitrary values of x. Accordingly, it is proved that CGD h is a probability density function.

Bibliography

[1] K. Shima *et al.*, "Measurement and Evaluation of Finger Tapping Movements Using Log-linearized Gaussian Mixture Networks," Sensors 2009, vol. 9(3), 2187-2201 (2009)

[2] H. Jiang *et al.*, "An Automatic Detection System of Lung Nodule Based on Multi-group Patch-Based Deep Learning Network," IEEE J. Biomed. Heal. Informatics, vol. 22, no. 4, pp. 1227-1237, 2018.

[3] K. Englehart *et al.*, "Classification of the myoelectric signal using time-frequency based representations", Medical Engineering & Phiysics, vol. 21, pp. 431-138 (1999)

[4] O. Fukuda *et al.*, "A human-assisting manipulator teleoperated by EMG signals and arm motions," IEEE Trans. on Roboticsand Automation, vol. 19-2, pp.210-222, 2003.

[5] T. Tsuji *et al.*, "A Recurrent Log-Linearized Gaussian Mixture Network," IEEE Trans. on Neural Networks, vol. 14-2, pp.304-316, 2003.

[6] A. Furui *et al.*, "A myoelectric prosthetic hand with muscle synergy-based motion determination and impedance model-based biomimetic control," Science Robotics, vol. 4, no. 31, eaaw6339, 2019.

[7] M. Akhbari *et al.*, "ECG segmentation and fiducial point extraction using multi hidden Markov model," Computers in Biology and Medicine, vol. 79, pp.21-29, 2016.

[8] P. Perera *et al.*, "Learning Deep Features for One-Class Classification," IEEE Trans. on Image Processing, vol. 28, Issue 11, pp. 5450-5463, 2019.

[9] K. Simonyan *et al.*, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, 1409.1556,, 2015.

[10] H. He *et al.*, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009,

[11] M. Luo *et al.*, "Using imbalanced triangle synthetic data for machine learning anomaly detection," Computers, Materials and Continua, vol. 58, no. 1, pp. 15-26, 2019.

[12] Zhi-Hua Zhou *et al.*, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 63-77, 2006.

[13] H. Masnadi-Shirazi *et al.*, "Cost-Sensitive Boosting," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 294-309, 2011.

[14] W. Lu *et al.*, "Feature fusion for imbalanced ECG data analysis," Biomedical Signal Processing and Control, vol. 41, pp. 152-160, 2018.

[15] M. Ribeiro *et al.*, "A study of deep convolutional auto-encoders for anomaly detection in videos," Pattern Recognition Letters, vol. 105, pp. 13-22, 2018.

[16] E. M. Knorr, *et al.*, "Distance-based outliers: algorithms and applications," The VLDB Journal, vol. 8, pp. 237-253, 2000.

[17] C. M. Rocco S *et al.*, "A support vector machine integrated system for the classification of operation anomalies in nuclear components and systems," Reliability Engineering and System Safety, vol. 92, Issue 5, pp. 593-600, 2007.

[18] J. Mourao-Miranda *et al.*, "Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine," Neuroimage, vol. 58, no. 3, pp. 793-804, 2011.

[19] F. Tony Liu *et al.*, "Isolation Forest," IEEE Int. Conf. on Data Mining, pp. 413-422, 2008.

[20] R. Chalapathy *et al.*, "Anomaly Detection using One-Class Neural Networks," *arXiv preprint arXiv:1802.06360*, 2018.

[21] K. Shima *et al.*, "A Novel Classification Method with Unlearned-class Detection Based on a Gaussian Mixture Model," IEEE Int. Conf. on Systems, Man, and Cybernetics, pp. 3726-3731, 2014.

[22] T. Tsuji *et al.*, "A Log-Linearized Gaussian Mixture Network and Its Application to EEG Pattern Classification," vol. 29, no. 1, pp. 60-72, 1999.

[23] S. Z. Yu, "Hidden semi-Markov models", Artificial Intelligence, vol. 174, Issue 2, pp. 215-243, 2010.

[24] S. Z. Yu and H. Kobayashi, "Practical Implementation of an Efficient Forward-backward Algorithm for an Explicit-duration Hidden Markov Model", IEEE Trans. on Signal Processing, Vol. 54, Issue 5, pp. 1947-1951, 2006.

[25] R. Domingues *et al.*, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," Pattern Recognition, vol. 74, pp. 406-421, 2018.

[26] J. S. Kim *et al.*, "Robust kernel density estimation," Journal of Machine Learning Research, vol. 13, pp. 2529-2565, 2012.

[27] M. E. Tipping *et al.*, "Probabilistic Principal Component Analysis," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 61, Issue 3, pp. 611-622, 1999.

[28] C. M. Bishop, "Pattern Recognition and Machine Learning ," Springer, 2006

[29] N. Nasios *et al.*, "Variational learning for Gaussian mixture models," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 4, pp. 849-862, 2006.

[30] Z. Zhang, *et al.*, "Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm," Statistics and Computing, no. 14, pp. 343-355, 2004.

[31] C. E. Rasmussen, "The infinite Gaussian mixture model," Advances in Neural Information Processing Systems, pp. 554-559, 2000.

[32] B. Bose *et al.*, "Detecting Insider Threats Using RADISH: A System for Real-Time Anomaly Detection in Heterogeneous Data Streams," IEEE Systems Journal, vol. 11, no. 2, pp. 471-482, 2017.

[33] M. Munir *et al.*, "DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series," IEEE Access, Vol. 7, pp. 1991-2005, 2019.

[34] A. A. Cook *et al.*, "Anomaly Detection for IoT Time-Series Data: A Survey," IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6481-6494, 2020.

[35] T. Fuse *et al.*, "Statistical Anomaly Detection in Human Dynamics Monitoring Using a Hierarchical Dirichlet Process Hidden Markov Model," IEEE Trans. on Intelligent Transportation Systems, vol. 18, no. 11, pp. 3083-3092, 2017.

[36] J. Bae *et al.*, "Gait phase analysis based on a Hidden Markov Model," Mechatronics, Vol. 21, Issue 6, pp. 961-970, 2011.

[37] T. Mukaeda *et al.*, "A novel hidden Markov model-based pattern discrimination method with the anomaly detection for EMG signals," Proceedings of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, pp. 921-924, 2017.

[38] J. Chorowski *et al.*, "Learning Understandable Neural Networks With Nonnegative Weight Constraints," IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 1, pp. 62-69, 2015.

[39] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE Int. Conf. on Computer Vision (ICCV), pp. 618-626, 2017.

[40] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, vol.77, no. 2, pp. 257-286, 1989.

[41] W. Zeng *et al.*, "Classification of neurodegenerative diseases using gait dynamics via deterministic learning", Information Sciences, vol.317, pp. 246-258, 2015.

[42] D. Y. Yeung, *et al.*, "Host-based intrusion detection using dynamic and static behavioral models", Pattern Recognition, vol. 36, no. 1, pp.229-243, 2003.

[43] J. M. Hausdorff *et al.*, "Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington's disease", J. Appl. Physiol., vol. 82, no. 1, pp. 262-269, 1997

[44] J.M. Hausdorff *et al.*, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," J. Appl. Physiol., vol. 88, no. 6, pp. 2045-2053, 2000.

[45] PhysioNet, "Gait Dynamics in Neuro-Degenerative Disease Data Base," from `http://www.physionet.org/physiobank/database/gaitndd`

[46] E. Baratin *et al.*, "Wavelet-based characterization of gait signal for neurological abnormalities," Gait Posture, vol. 41, no. 2, pp. 634-639, 2015.

[47] Q. Ye *et al.*, "Classification of Gait Patterns in Patients with Neurodegenerative Disease Using Adaptive Neuro-Fuzzy Inference System," Computational and Mathematical Methods in Medicine, vol. 2018, 2018.

[48] S. A. A. Shah *et al.*, "Classification of Control and Neurodegenerative Disease Subjects Using Tree Based Classifiers," J. Pharm. Res. Int., vol. 32, no. 11, pp. 63-73, 2020.

[49] S. Miyajima *et al.*, "Minimal Inertial Sensor Placement for Work Recognition and Working Posture Assessment," 2019 IEEE/SICE International Symposium on System Integration (SII), pp. 391-395, 2019.

[50] Z. Cao *et al.*, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2021.

[51] H. P. H. Shum *et al.*, "Real-Time Posture Reconstruction for Microsoft Kinect," in IEEE Transactions on Cybernetics, vol. 43, no. 5, pp. 1357-1369, 2013.

[52] J. Duchi *et al.*, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," Journal of Machine Learning Research, vol. 12, pp. 2121-2159, 2011.

[53] M. Zak, "Terminal attractors in neural networks," Neural Networks, vol. 2, Issue 4, pp. 259-274, 1989.

[54] C. He *et al.*, "A Fully-Pipelined Hardware Design for Gaussian Mixture Models," IEEE Transactions on Computers, vol. 66, no. 11, pp. 1837-1850, 2017.

[55] N. Bu *et al.*, "FPGA implementation of a probabilistic neural network for a bio-electric human interface," Midwest Symposium on Circuits and Systems, vol. 3, no. 2, pp. 3-6, 2004.

[56] D. Gorur *et al.*, "Dirichlet process gaussian mixture models: Choice of the base distribution," Journal of Computer Science and Technology, vol. 25, no. 4, pp. 653-664, 2010.

[57] M. J. Beal *et al.*, "The Infinite Hidden Markov Model," Advances in Neural Information Processing Systems, pp. 577-584, 2002.

Published Papers

Refereed papers

- <u>Takayuki Mukaeda</u> and Keisuke Shima, "A HMM-based pattern recognition method with unlearned pattern detection and its application to biosignal classifications (in Japanese)", Transactions of the Society of Instrument and Control Engineers, vol. 54, no. 1, pp. 9-15, 2018.

- <u>Takayuki Mukaeda</u> and Keisuke Shima, "A novel probabilistic neural network with unlearned-class detection based on normal and complementary Gaussian mixture models (in Japanese)", Transactions of the Society of Instrument and Control Engineers, vol. 56, no. 12, pp. 532-540, 2020.

- <u>Takayuki Mukaeda</u> and Keisuke Shima, "The one-vs-rest hidden Markov model-based pattern discrimination method with anomaly identification," IEEE Transactions on Pattern Analysis and Machine Intelligence (Under review)

Reviewed International Conference

- <u>Takayuki Mukaeda</u> and Nan Bu, "Forearm Motion Classification based on Frequency Characteristics of EMG Signals", 4th International Symposium on Technology for Sustainability 2014 (ISTS 2014), 2014.

- <u>Takayuki Mukaeda</u> and Nan Bu, "Wavelet feature selection using independent component analysis for forearm motions discrimination", Proc. of the Twenty-First International Symposium on Artificial Life and Robotics 2016 (AROB 21st 2016), pp. 865-870, 2016.

- Takayuki Mukaeda and Keisuke Shima, "A novel hidden Markov model-based pattern discrimination method with the anomaly detection for EMG signals", 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '17), pp. 921-924, 2017.

- Takayuki Mukaeda, Keisuke Shima, Saori Miyajima, Yuki Hashimoto, Takayuki Tanaka, Naomichi Tani and Hiroyuki Izumi, "Development of an anomaly detection method with a novel hidden semi-Markov model incorporating unlearned states", 2020 IEEE/SICE International Symposium on System Integration (SII), pp. 1270-1275, 2020.

Reviewed International Conference(co-author)

- Nan Bu and Takayuki Mukaeda, "Classification of Simultaneous Motions using Wavelet-based Time-frequency Representation of EMG Signals", The 4th IIAE International Conference on Intelligent Systems and Image Processing 2016 (ICISIP2016), pp. 465-470, 2016.

- Ryoma Mitani, Koji Shimatani, Mami Sakata, Takayuki Mukaeda and Keisuke Shima, "Effects of somatosensory information provision to fingertips for mitigation of postural sway and promotion of muscle coactivation in an upright posture", The 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '19), pp. 5096-5099, 2019.

Reviewed Domestic Conference

- Takayuki Mukaeda and Keisuke Shima, "A novel probabilistic neural network incorporating complementary Gaussian distribution with unlearned-class detection (in Japanese)", 25th Robotics symposia, 2020

Domestic Conference

- 迎田 隆幸, 卜 楠, "Wavelet 変換を利用した前腕筋電動作識別", 日本人間工学会第 35 回九州・沖縄支部大会, pp. 33-37, 2014.

- 迎田 隆幸, 島 圭介, "未学習クラス推定を可能とする HMM の提案と時系列筋電パターンの識別", ロボティクス・メカトロニクス講演会 2017, 2A2-J03, 2017

- 迎田 隆幸, 島 圭介, 宮島 沙織, 橋元 裕紀, 田中 孝之, 谷 直道, 泉 博之, "未分類状態を考慮した隠れセミマルコフモデルによる時系列パターン解析", 第 28 回インテリジェント・システム・シンポジウム (FAN 2018), 2018.

- 島 圭介, 迎田 隆幸, 田中 孝之, 宮島 沙織, 橋元 裕紀, 谷 直道, 泉 博之, "ウェアラブル慣性センサと隠れセミマルコフモデルに基づく介護記録自動化システム", 第 20 回計測自動制御学会システムインテグレーション部門講演会講演論文集, pp.1180-1182, 2019.

- 井澤 啓太, 迎田 隆幸, 島 圭介, "線形化隠れセミマルコフモデルに基づくリカレントニューラルネット", 計測自動制御学会システム・情報部門学術講演会 2019 講演論文集, pp. 576-578, 2019.

- 清水 武史, 迎田 隆幸, 島 圭介, 村治 周, 松尾 純太郎, 堀上 正義, "多次元生体信号解析に基づくドライバの精神的負荷リアルタイム推定法", 日本人間工学会第 60 回大会講演集, 2C3-3, 2019.

- 堀内 太貴, 迎田 隆幸, 島 圭介, 島谷 康司, "モバイル距離画像センサを用いた創傷の定量評価システム", 第 20 回計測自動制御学会システムインテグレーション部門講演会講演論文集, pp.520-522, 2019.

- 照井 洸貴, 迎田 隆幸, 内山 政哉, 島 圭介, 島谷 康司, "肢体不自由者のための FES 型 BCI トレーニング法", 第 20 回計測自動制御学会システムインテグレーション部門講演会講演論文集, pp.696-697, 2019.

- 植草 秀明, 迎田 隆幸, 清水 武史, 島 圭介, "FPGA に実装を指向した近似 GMM に基づく未学習クラス推定ニューラルネット", ロボティクス・メカトロニクス講演会 2020, 1P2-F05, 2020.