Novelty Detection in Multimodal Datasets Based on Least Square Probabilistic Analysis

Hiroyuki Yoda, Akira Imakura, Momo Matsuda, Xiucai Ye, and Tetsuya Sakurai

***Abstract*—Novelty detection represents the detection of anomalous data based on a training set consisting of only the normal data. In this study, we propose a new probabilistic approach for novelty detection to effectively detect anomalous data, particularly for the case of multimodal training dataset. Our method is inspired by the Least-Squares Probabilistic Classifier (LSPC), which is an efficient multi-class classification method. Numerical experimental results based on multimodal datasets show that the proposed method outperforms the related methods.**

***Index Terms*—Novelty detection, multimodal datasets, least-square probabilistic analysis.**

I. Introduction

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [1]. These nonconforming data are most commonly referred to as anomalies or outliers [2], [3]. In anomaly detection, to detect anomalies without having any previous knowledge of their characteristics, the outlier detection and novelty detection are frequently employed. The difference between these two methods is that the outlier detection is employed when the anomalies are expected to be part of the training data, while the novelty detection is employed when all training data is normal. Thus, novelty detection is the process of identifying the observations that differ in some respect from the ones present in the training set [4].

One of the main obstacles in the field of anomaly detection is that it is difficult to collect sufficient anomalous data for training. In contrast, collecting sufficient normal data is an easy task. Consequently, novelty detection has been extensively applied in many research areas. Examples include medical diagnostic problems [5], failure detection in complex industrial problems [6], sensor networks [7], video surveillance [8], and detection of masses in mammograms [9]. Owing to its applicability and importance, various methods of novelty detection have been studied such as probabilistic, distance-based, domain-based, reconstruction-based, and information theoretic method [3].

In this study, we focus on the probabilistic method due to its computational efficiency. Probabilistic approaches to novelty detection are divided into two steps. First, during training, the probabilistic density function of normal data is measured. Then, the test data is classified as an anomaly if its probabilistic density value is low.

Manuscript received September 15, 2019; revised March 22, 2020.

The authors are with University of Tsukuba, Ibaraki, Japan (e-mail: yoda.hiroyuki.wy@alumni.tsukuba.ac.jp, imakura@cs.tsukuba.ac.jp, matsuda.momo.ww@alumni.tsukuba.ac.jp, yexiucai2013@gmail.com, sakurai@cs.tsukuba.ac.jp).

There are two kinds of methods applied to estimate probabilistic density functions, namely parametric and non-parametric methods. The parametric method assumes that the data follows a certain distribution. In contrast, the non-parametric model does not expect that the data to follow a specific distribution. Since there are very few data follow a certain distribution, we apply the non-parametric method considering its multiplicity of use.

A non-parametric method for the probability density estimation is the Kernel Density Estimation (KDE) [10]. KDE is a flexible approach used to estimate the densities of given data distribution, that contains no information on the underlying distribution [11]. However, when the training data is multimodal, using KDE for novelty detection will decrease the classification performance. The case where the normal data is multimodal means that there are two or more points with high density in normal data. In this study, multimodal data is defined as the case that the normal data includes two or more classes.

Therefore, by applying KDE to multimodal data, probabilistic density function tends to classify the points with high density as the anomalous data. This is because data points in the middle of multiple classes, which do not belong to any of them, are influenced by each class, resulting in an unnecessarily high score, which leads to the misclassification of anomalous data as normal.

In our study, we propose a novelty detection method based on least-square probabilistic analysis. The proposed method is inspired by Least-Squares Probabilistic Classifier (LSPC) [12], which is a machine learning algorithm that performs multi-class classification by focusing on the ratio of the data density of each class and the density of all data.

Our method is an effective method for detecting anomalies when the normal data is multimodal because it calculates the novelty score only to the closest class, but also not influenced by each class. In the case when the training data does not have class information, we assign class information to preprocess the training data by using X-means [13], which is a clustering method used when there is no information on the number of classes by determining the number of classes and applying k-Means clustering.

Our contributions of this study are summarized as follows.

- We propose a novelty detection method based on least-squares probabilistic analysis, which is applicable for multimodal datasets.
- Using X-means clustering, we enable the application of proposed methods when the training data has no class information.
- Our method shows competitive results in both artificial dataset and benchmark datasets. In the experiments, the

doi: 10.18178/ijmlc.2020.10.4.968

proposed method achieves a very good result in the USPS dataset and even better results regarding mislabeled detection (more information on mislabeled detection is provided in Ref. [13]).

II. Preliminaries

A. Notation

Suppose that we are given a training set of $n \in \mathbb{N}$ samples

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$$

drawn independently from a joint probability distribution with density $p(\boldsymbol{x}, y)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the d-dimensional feature vector,

$$y_i \in \{0,1,\dots,Y\}$$

is a class label, and $Y \in \mathbb{N}$ is the number of normal classes. Note that here, we define class label "0" as the "anomalous" class and the other as the "normal" class, while there are no data belong to class "0" in the training data.

The objective of novelty detection is to distinguish whether test data $\boldsymbol{x}^{\mathrm{te}} \in \mathbb{R}^d$ is normal or anomalous by using the training set.

B. Kernel Density Estimation

KDE estimates the density $p(\boldsymbol{x})$ from the training data $\boldsymbol{x}_i \in \mathbb{R}^d (i = 1,2,\dots,n)$ by the function below.

$$p(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{n\sigma}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathbf{exp}\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x_i}\|_2^2}{\boldsymbol{2\sigma^2}}\right), \tag{1}$$

where $\sigma \in \mathbb{R}$ denotes the bandwidth.

From Equation (1), the value of $p(\boldsymbol{x})$ increases if there is a large amount of training data distributed around the input data, whereas it assumes a low value if there are few training data in the vicinity.

Since we only use normal data for training, the density $p(\boldsymbol{x})$ can be defined as a normality score. This is because $p(\boldsymbol{x})$ tends to provide high-density values for regions with a lot of normal data and low-density values for regions with few normal data. To transform the normality score to the novelty score, we define the novelty score $w(\boldsymbol{x}^{\mathrm{te}})$ as

$$w(\boldsymbol{x}^{\mathrm{te}}) = t - p(\boldsymbol{x}^{\mathrm{te}}),$$

where $t \in \mathbb{R}$ is the maximum value of $p(\boldsymbol{x}^{\mathrm{te}})$ for all test data $\boldsymbol{x}^{\mathrm{te}} \in \mathbb{R}^d$.

By using novelty score $w(\boldsymbol{x}^{\mathrm{te}})$, test data can be classified by a predetermined threshold $\tau \in \mathbb{R}$, such that test data $\boldsymbol{x}^{te}$ is classified as normal if the novelty score $w(\boldsymbol{x}^{\mathrm{te}})$ is "smaller" than τ, and anomalous if the novelty score $w(\boldsymbol{x}^{\mathrm{te}})$ is "larger" than τ. Note that here, we define data as anomalous if the novelty score is bigger than the threshold while Sugiyama [14] considers the opposite.

III. Proposed Method

In this section, we propose a novelty detection method (ND-LSPA) in multimodal datasets based on least-square probabilistic analysis.

Let $p(y|\boldsymbol{x})$ be the class-posterior probability. The objective of ND-LSPA is to learn the novelty score $w(\boldsymbol{x})$ by using the training set

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}.$$

Since class label "0" is the anomaly class, the novelty score $w(\boldsymbol{x})$ is represented as

$$w(\boldsymbol{x}) \cong p(0|\boldsymbol{x}).$$

Based on the idea of LSPC, we describe how to learn the class-posterior probability $p(y|\boldsymbol{x})$ from the training set.

For each $y \in \{0,1,\dots,Y\}$, we model $p(y|\boldsymbol{x})$ by

$$q(y|\boldsymbol{x};\boldsymbol{\alpha}_y) := \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}),$$

where

$$\boldsymbol{\alpha}_y = (\alpha_{y,0}, \alpha_{y,1}, \dots, \alpha_{y,n})^{\mathrm{T}} \in \mathbb{R}^n,$$

is the weight vector for each class,

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x_1}), \dots, \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x_n}))^{\mathbf{T}} \in \mathbb{R}^{\boldsymbol{n}}, \tag{2}$$

is the basis function vector, and $k(\boldsymbol{x}, \boldsymbol{x}')$ is the kernel function.

Using the model $q(y|\boldsymbol{x};\boldsymbol{\alpha}_y)$, we determine the weight vector $\boldsymbol{\alpha}_y$ which minimize squared loss J defined below.

$$J(\boldsymbol{\alpha}_y) := \frac{1}{2}\int (q(y|\boldsymbol{x};\boldsymbol{\alpha}_y) - p(y|\boldsymbol{x}))^2 \, p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where $p(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$ and assume that $p(\boldsymbol{x})$ is positive for all $\boldsymbol{x} \in \mathbb{R}^d$. Expanding the squared term, we can express J as

$$J(\boldsymbol{\alpha}_y) = \frac{1}{2}\int \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\alpha}_y p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) p(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} + \mathrm{Const},$$

where $p(y|\boldsymbol{x}) = p(\boldsymbol{x}|y)p(y)/p(\boldsymbol{x})$ is used.

By taking the sample mean for class $y \in \{1,\dots,Y\}$, the first and second terms can be approximated as

$$\int \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\alpha}_y p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\alpha}_y,$$

$$\int \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) p(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{i:y_i=y} \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i).$$

Thus, to compute the weight vector $\boldsymbol{\alpha}_y$, by ignoring the constant, we minimize the approximated squared loss with regularization term

$$\frac{1}{2n}\sum_{i=1}^{n} \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\alpha}_y - \frac{1}{n} \sum_{i:y_i=y} \boldsymbol{\alpha}_y^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\alpha}_y\|^2,$$

where $\lambda \in \mathbb{R}$ is the regularization parameter. Taking the derivative, the weight vector $\boldsymbol{\alpha}_y$ can be obtained by solving the following linear system

$$(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda n I_n)\boldsymbol{\alpha}_y = \boldsymbol{\Phi}^{\mathrm{T}}[\delta_{y,y_1}, \dots, \delta_{y,y_n}]^{\mathrm{T}}, \tag{3}$$

where

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}(\boldsymbol{x}_1), \dots, \boldsymbol{\phi}(\boldsymbol{x}_n)]^{\mathrm{T}} \in \mathbb{R}^{n \times n}, \tag{4}$$

I_n is the identity matrix and $\delta_{y,y_i} \in \{0,1\}$ is the Kronecker's delta defined as

$$\delta_{y,y_i} = \begin{cases} 1 & (y_i = y), \\ 0 & (y_i \neq y). \end{cases} \quad (5)$$

Let $\widehat{\boldsymbol{\alpha}}_y$ be the solution of Equation (3), the class-posterior probabilities $q(y|\boldsymbol{x};\boldsymbol{\alpha}_y)$ can be estimated as

$$\hat{q}(y|\boldsymbol{x}) = q(y|x;\widehat{\boldsymbol{\alpha}}_y) = \widehat{\boldsymbol{\alpha}}_y{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) \quad (6)$$

for $y \in \{1, \ldots, Y\}$.

For $y = 0$, i.e., the anomalous class, since there is no training data, we cannot use the same model as $y \in \{1, \ldots, Y\}$. Here, from the definition of the class-posterior probability $p(0|\boldsymbol{x})$, we have

Algorithm 1 Novelty Detection based on Least Square Probabilistic Analysis

Input: Training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, test sample $\boldsymbol{x}^{\mathrm{te}}$, bandwidth σ for the kernel function, and regularization parameter λ.
Output: Novelty score $w(\boldsymbol{x}^{\mathrm{te}})$
1: Set the matrix $\boldsymbol{\Phi}$ (4)
2: Solve the linear systems (3) for $y = 1, \ldots, Y$
3: Set class posterior probability $\hat{q}(y|\boldsymbol{x}) = \boldsymbol{\alpha}_y^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$ for $y = 1, \ldots, Y$
4: Calculate $\tilde{q}(y|\boldsymbol{x})$ by Eq. (8)
5: Calculate $w(\boldsymbol{x}^{\mathrm{te}}) = \hat{q}(0|\boldsymbol{x}^{\mathrm{te}})$ by Eq. (7)

$$p(0|\boldsymbol{x}) = 1 - \sum_{y=1}^{Y} p(y|\boldsymbol{x}).$$

Then, we model $\hat{q}(0|\boldsymbol{x})$ as

$$\hat{q}(0|\boldsymbol{x}) := 1 - \sum_{y=1}^{Y} \tilde{q}(y|\boldsymbol{x}), \quad (7)$$

where

$$\tilde{q}(y|\boldsymbol{x}) = \begin{cases} \dfrac{\max(0, \hat{q}(y|\boldsymbol{x}))}{\rho} & \left(\begin{array}{l}\hat{q}(y|\boldsymbol{x}) \geq \hat{q}(y_i|\boldsymbol{x}) \\ \text{for all } y_i \in \{1, \ldots, Y\}\end{array}\right) \\ 0 & (\text{otherwise}) \end{cases}, \quad (8)$$

and

$$\rho = \max_{\substack{\boldsymbol{x} \in \{\boldsymbol{x}^{\mathrm{te}}, \boldsymbol{x}_i (i=1,\ldots,n)\}, \\ y_i \in \{1,\ldots,Y\}}} \hat{q}(y|\boldsymbol{x}).$$

Finally, we obtain the novelty score $w(\boldsymbol{x})$ as

$$w(\boldsymbol{x}) = \hat{q}(0|\boldsymbol{x}),$$

that satisfies $0 \leq w(\boldsymbol{x}) = \hat{q}(0|\boldsymbol{x}) \leq 1$.We summarize the procedure of the proposed method in Algorithm 1.

Note that here, this algorithm assumes that there is a class label information associated with each training data. In the case when there is no class information associated, we employ X-means to training data in order to assign class information.

IV. Experimental Results

In this section, the proposed method (ND-LSPA) is evaluated on an artificial dataset and several benchmark datasets. The performance of the ND-LSPA is compared with KDE and the Kullback-Liebler Importance Estimation Procedure (KLIEP) [15] proposed by Sugiyama.

In this study, we employed the kernel function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\sigma^2}\right),$$

as the base function written in Equation (2), where $\sigma \in \mathbb{R}$ is the bandwidth of the kernel function.

To determine the bandwidth $\sigma \in \mathbb{R}$ for KDE, we applied the Silverman's rule [16]. For ND-LSPA, to determine

bandwidth $\sigma \in \mathbb{R}$, we replicated the procedure from the Zelnik-Manor and Perona [17]. More information is provided in their work. Furthermore, the regularization parameter $\lambda \in \mathbb{R}$ is set as 0.01.

We used Area Under ROC Curve (AUC) [18] to evaluate the classification performance, which enables to calculate the performance without the predetermined threshold $\tau \in \mathbb{R}$. AUC assumes values from 0 to 1, and a higher classification performance brings the value closer to 1.

TABLE I: Breakdown of Artificial Datasets

Dataset	Normal Data (Training Data)		Anomalies
	Class 1	Class 2	
1	0~0.3	0.7~1.0	0.3~0.7
2	0~0.4	0.6~1.0	0.4~0.6

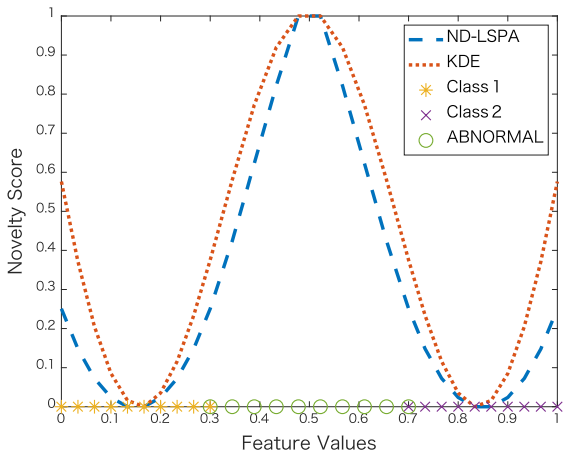


Fig. 1. Result for artificial dataset 1.

A. Artificial Data

We conduct an experiment using a one-dimensional artificial dataset to examine the effects of the different methods. The goal of this experiment is to verify if ND-LSPA performs better than KDE when the training datasets are composed of two normal classes and the interval between them is small. We highlight that data in these intervals is considered to be anomalous.

We compare the results by visualizing the novelty score function values for each method and we also evaluate the classification performance using AUC values.

1) Datasets

We used two artificial datasets consisting of 20 training examples and 30 test samples. To achieve multimodality in the training datasets, the training datasets are composed of two classes with completely different intervals.

For the first artificial dataset used for training, the data for class 1 is obtained by dividing the [0, 0.3] interval equally into 10 data points. The data for class 2 uses the interval of [0.7 1.0]. For the second artificial dataset used for training,

the data for class 1 is obtained by dividing the [0, 0.4] interval into equally 10 data points. The data for class 2 involves the interval of [0.6, 1.0].

Here, we divide the [0, 1] interval equally into 30 data points for the test data. We consider that examples in the interval [0.3, 0.7] are anomalies for the first test data. For the second test data, we apply the same method, but we limit the interval to [0.4, 0.6]. The breakdown of the artificial dataset is shown in Table I.

2) Results of artificial data

Using the datasets shown in Table I, we visualized the function of the novelty score $w(x)$ for each method and compared its performance using AUC values. Notice that here, we applied min-max transformation to rescale the novelty score of KDE to [0,1] in order to compare easily. From the results in Fig. 1, we can see that the novelty scores of ND-LSPA and KDE are higher values for the anomalies.

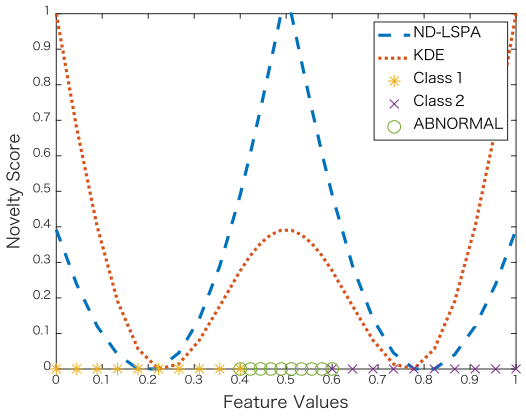


Fig. 2. Result for artificial dataset 2.

The AUC value was 0.97 for the KDE, and 0.99 for the ND-LSPA. Thus, both methods were able to detect anomalies with high accuracy even if the training data has multimodality.

In contrast, the results in Fig. 2 indicate that the novelty score of KDE gives a low value in the anomalous points, while ND-LSPA maintains a high value. Furthermore, the AUC values for the second dataset was 0.69 for KDE and 0.99 for ND-LSPA. Therefore, from the results of this experiment, the classification performance of the KDE decreases while the ND-LSPA maintains high performance when the interval of anomalies becomes smaller in our artificial multimodal datasets.

B. MNIST Datasets

In this section, we compare the proposed method with KDE and KLIEP using the MNIST datasets especially when the multimodality, in other words, the number of classes in training dataset increases.

1) Datasets

MNIST [19] is a dataset of handwritten numeric images from "0" to "9," consisting of a training set consisting of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 pixel grayscale image, associated with a label from 10 classes.

2) Setup

To confirm the effectiveness of the proposed method for multimodal data, we recombine the classes of the MNIST dataset into two groups, namely a normal group and an anomalous group. We repeat this recombination nine times. At first, the normal group only has one class, whereas the anomalous group has nine classes. We call this "dataset-1" (because there is one class in the normal group). Then, for "dataset-2," we have two classes in the normal group (amounting to eight in the anomalous group). For "dataset-3" through "dataset-9" we maintain this process of increasing the normal group with one class and correspondingly decreasing the anomalous group by one class. For example, "dataset-9" has nine classes in the normal group (i.e., one class in the anomalous group). We randomly choose which class is to be normal (and anomalous) among "0" to "9" digits (classes). Each dataset consists of a training set of 1,000 examples and a test set of 500 examples. The test set includes 475 normal data and 25 anomalous data. Since the performance changes using different digits (classes) for training, we repeat the experiment 100 times. In addition, we preprocess all images with a min-max transformation to rescale the data to [0,1]. Subsequently, we reduce the dimensionality of the data via PCA and we chose the minimum number of eigenvectors, such that at least 80% of the variance is retained.

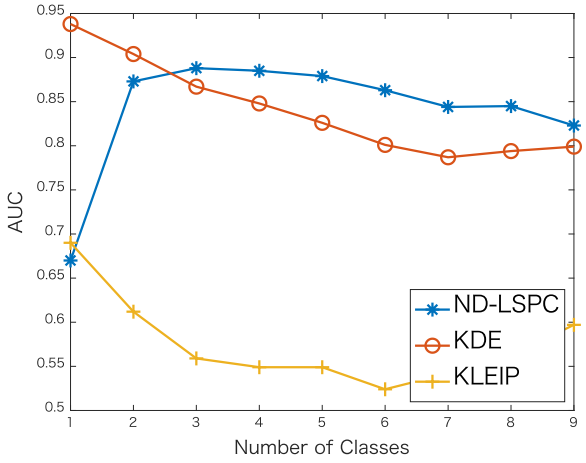


Fig. 3. AUC results vs. number of classes.

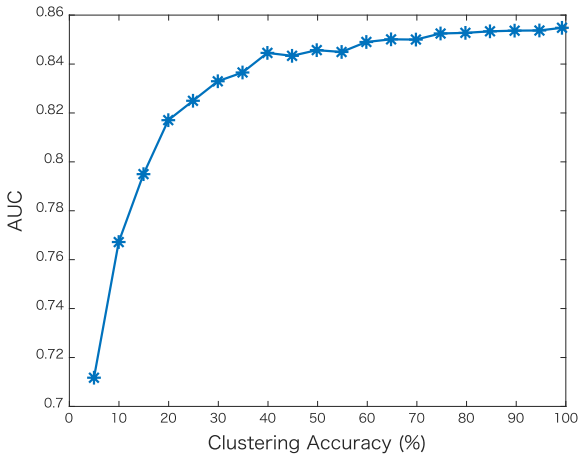


Fig. 4. Relation between clustering accuracy and AUC.

3) Results of MNIST datasets

As shown in Fig. 3, the classification performance of KDE and KLIEP decreases as the number of normal classes, hence the multimodality of training data increases. In contrast, the proposed method maintains higher

classification performance when the multimodality of training datasets increases. Therefore, our method demonstrates better classification performance for multimodal datasets.

C. Clustering Accuracy and AUC

We show the relationship between clustering accuracy and AUC using ND-LSPA, in case there is no class label information associated with the training data. In this study, we used NMI [20] to calculate the clustering accuracy.

1) Setup

To conduct the experiments, we arbitrarily chose "dataset-5", described in Section 4.2.2. Since we wish to verify the relationship between AUC values and clustering accuracy, we methodologically changed some class labels leading to a mislabel situation. Hence, with no changes, the labels match the features perfectly, indicating that the clustering accuracy is 100%. With changes, the clustering accuracy decreases.

We report the results when the label changes lead to a clustering accuracy decrease of about 5%. Thus, we show the AUC values when the clustering accuracy reaches approximately [5%, 10%, …, 95%, 100%]. For each accuracy level, we repeat the experiments 100 times and calculate the mean value of AUC.

TABLE II: AVERAGE AUCS WITH STANDARD DEVIATIONS BY DIFFERENT METHODS FOR BENCHMARK DATASETS

Dataset	KDE	KLIEP	ND-LSPA	ND-LSPA X-Means
MNIST	0.826 ± 0.08	0.549 ± 0.10	**0.879 ± 0.05**	0.780 ± 0.08
Fashion-MNIST	0.734 ± 0.07	0.619 ± 0.12	0.736 ± 0.09	**0.756 ± 0.08**
EMNIST	**0.918 ± 0.03**	0.665 ± 0.07	0.803 ± 0.05	0.694 ± 0.05
USPS	0.835 ± 0.10	0.591 ± 0.13	**0.901 ± 0.03**	0.822 ± 0.06

2) Results of clustering accuracy and AUC

Fig. 4 shows that the AUC value tends to increase with increasing the clustering accuracy. Furthermore, the correlation coefficient was 0.73, indicating a positive correlation between two indicators.

D. Benchmark Dataset

In this section, we compare the proposed method with other related methods by using benchmark datasets in two conditions, including the use of class information and no use of the class information for training. For the training datasets with no class information, we employed X-means to obtain class information.

1) Datasets

Fashion-MNIST [21]: The Fashion-MNIST dataset shares the same image size, class size and the structure of training and testing splits with the MNIST dataset. The difference is that the images are consist of 10 types of clothes.

EMNIST [22]: EMNIST is a dataset of handwritten digits "0" to "9" and handwritten alphabetical letters, which consists of a training set of 60,000 examples and test set of 10,000 examples for handwritten digits and 124,800 examples of handwritten letters. The images' size is the same as that of MNIST and Fashion-MNIST datasets.

USPS [23]: USPS is a dataset of handwritten numeric images from "0" to "9" that comprised a training set of 7,291 examples and a test set of 2,007 examples. Each example is a 16×16 grayscale image, associated with a label from 10 classes.

2) Setup

For MNIST, Fashion-MNIST, USPS, we randomly chose five classes among 10 to be normal, and the other five classes to be anomalous. For EMNIST, we assume that all handwritten digits "0" through "9" are normal, and all the alphabet handwritten alphabetical letters are anomalous data.

The other settings such as the number of samples for training data and test data, repetition time, and the number of features are the same as in Section IV.B.2.

3) Results of benchmark datasets

The results shown in Table II indicate that ND-LSPA with using clustering information shows higher AUC values of MNIST, Fashion-MNIST, and USPS. In contrast, the EMNIST dataset demonstrates higher AUC values with KDE than ND-LSPA. The reason for this result might due to the difference in the number of classes included in the training data. While other datasets include five classes in the training data, the EMNIST dataset includes 10 classes in the training data with the same number of examples as the other datasets. This made it difficult to sufficiently train for each class due to the lack of each class samples, and led to lower classification performance.

TABLE III: AVERAGE AUCS WITH STANDARD DEVIATIONS BY DIFFERENT METHODS FOR MISLABELED DATASETS

Dataset	KDE	KLIEP	ND-LSPA
MNIST	0.828 ± 0.04	0.705 ± 0.14	**0.936 ± 0.02**
Fashion-MNIST	0.899 ± 0.03	0.781 ± 0.13	**0.938 ± 0.03**
EMNIST	0.859 ± 0.03	0.669 ± 0.07	**0.875 ± 0.03**
USPS	0.877 ± 0.03	0.491 ± 0.07	**0.954 ± 0.02**

Regarding the result of the proposed method with no class information, although the AUC value increased for the Fashion-MNIST dataset, it decreased for the other dataset. However, as we show in Fig. 4 in Section IV.D, these results can be improved by using a better clustering method.

E. Mislabeled Detection

We compare the performance with the proposed method and related methods for mislabeled detection. To this end, we define the mislabeled detection as a method to find the incorrect class label in the test data.

1) Setup

The number of samples for training data, test data, and repetition time are the same as in Section IV.2. For MNIST, Fashion-MNIST, and USPS datasets, we used all 10 class labels for training data and test data. For the EMNIST dataset, we used 26 classes that refer to the handwritten alphabet.

To generate mislabeled data, we randomly selected 25 data from the test data and changed the class label.

Furthermore, we add 10 extra dimensions for MNIST, Fashion-MNIST, USPS and 26 extra dimensions to EMNIST corresponding to the class label. In the extra dimension, we add dummy variables with the size of the feature dimension instead of 1.

2) Results of mislabeled detection

Table III shows that the proposed method exhibits higher classification performance, especially for MNIST and USPS datasets. Comparing the AUC values of Fashion-MNIST and EMNIST datasets, ND-LSPA performs slightly better than KDE. However, since the standard deviation is small, it can be said that our proposed method can be considered better suited for mislabeled detection.

V. Conclusions

In this study, we proposed a novelty detection method in multimodal datasets based on least-square probabilistic analysis. Performing numerical experiments, we confirmed that the classification performance was higher than that in related methods, especially for MNIST and USPS. Furthermore, the proposed method provides a higher classification performance for mislabeled detection.

Future studies will consider improving the performance of our method by adjusting parameters such as the regularization parameter $\lambda \in \mathbb{R}$ and the bandwidth $\sigma \in \mathbb{R}$. Furthermore, we aim to extend this study by conducting experiments on real-world data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors conducted the research; Hiroyuki Yoda analyzed the data and wrote the paper; all authors had approved the final version.

Acknowledgements

The present study is supported in part by the Japan Science and Technology Agency (JST), ACT-I (No. JPMJPR16U6), the New Energy and Industrial Technology Development Organization (NEDO) and the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (Nos. 17K12690, 18H03250).

References

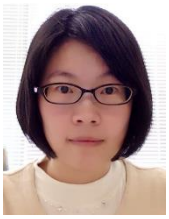
[1] V. Chandola, R. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, 2009.
[2] M. Pimentel, D. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, Institute of Biomedical Engineering Department of Engineering Science, University of Oxford, Oxford OX3 7DQ UK, vol. 99, pp. 215-249, 2014.
[3] Y. Xiucai and T. Sakurai, "Robust similarity measure for spectral clustering based on shared neighbors," *ETRI Journal*, pp. 540-550, 2016.
[4] S. Mohammad, K. Mohammad, F. Mahhamad, and A. Ehsan, "Adversarially learned one-class classifier for novelty detection," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3379-3388.
[5] L. Clifton, D. Clifton, P. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Proc. FedCSIS*, 2011, pp. 125-131.
[6] L. Tarassenko, D. Clifton, P. Bannister, and S. King, *Novelty Detection*, John Wiley & Sons, Ltd. pp. 1-22, 2010.
[7] Y. Zang, N. Merantnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Comm. Surv*, vol. 12, pp. 159-170, 2010.
[8] C. Diehl and J. Hampshire, "Real-time object classification and novelty detection for collaborative video surveillance," in *Proc. the International Joint Conference on Neural Networks*, 2002, vol. 3, pp. 2620-2625.
[9] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Proc. the 4th International Conference on Artificial Neural Networks*, IET, 1995, pp. 442-447.
[10] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp.1065-1076, 1962.
[11] F. Olga, Z. Enrico, and W. Ulrich, "Novelty detection by multivariate kernel density estimation and growing neural gas algorithm," *Mechanical Systems and Signal Processing*, vol. 50, pp. 427-436, 2015.
[12] M. Sugiyama, "Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting," *IEICE Transactions on Information and Systems*, vol. 93, pp. 2690-2701, 2010.
[13] D. Pelleg and A. Morre, "X-means: Extending K-means with efficient estimation of the number of clusters," Carnegie Mellon University, Pittsburgh, vol. 1, pp. 727-734, 2000.
[14] X. Kang, P. Duan, S. Li, and J. Benediktsson, "Detection and correction of mislabeled training samples for hyperspectral image classification," *Article in IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 5673-5686, 2018.
[15] S. Hido, Y. Tsuboi, H. Kashima, and M. Sugiyama, "Novelty detection by density ration estimation," *IBIS*, Tokyo, Japan, (November 5-7), 2007.
[16] W. S. Bernard, *Density Estimation for Statistics and Data Analysis*, School of Mathematics University of Bath, UK, 1986.
[17] L. Z. Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. NIP*, 2005, pp. 1601-1608.
[18] A. Bradley, "The use of the are under ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145-1159, 1997.
[19] H. Lin and C. Lin. *The MNIST Database of Handwritten Digits*. [Online]. Available: http://yann,lecen.com/exdb/mnist/
[20] N. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837-2854, 2010.
[21] S. Zaland. (2007). [Online]. Available: https://github.com/zalandoresearch/fashion-mnist/blob/master/README.ja.md
[22] G. Cohen, S. Afshar, J. Tapson, and A. Schaik. (2017). EMNIST: An extension of MNIST to handwritten letters. http://arxiv.org/abs/1702.05373
[23] Kaggle. [Online]. Available: https://www.kaggle.com/bistaumanga/usps-dataset

Hiroyuki Yoda is currently pursuing his master degree. He is now at the Department of Computer Science, University of Tsukuba, Japan. His current research interests include machine learning and anomaly detection.

Akira Imakura is an associate professor at the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. He received the Ph.D. in 2011 from Nagoya University, Japan. He was appointed as Japan Society for the Promotion of Science Research Fellowship for Doctor Course Student (DC2) from 2010 to 2011, as a research fellow at Center for Computational Sciences, University of Tsukuba, Japan from 2011 to 2013, and

also as a JST ACT-I researcher from 2016 to 2019. His current research interests include developments and analysis of highly parallel algorithms for large matrix computations. Recently, he also investigates matrix factorization-based machine learning algorithms. He is a member of JSIAM, IPSJ and SIAM.

Momo Matsuda is currently pursuing her doctor degree. She is now at the Department of Computer Science, University of Tsukuba, Japan. Her current research interests include dimensionality reduction, machine learning and spectral clustering.



Xiucai Ye is an assistant professor at the Department of Computer Science, University of Tsukuba, Tsukuba Science City, Japan. She received her PhD in computer science from University of Tsukuba, Tsukuba Science City, Japan, in 2014. Her current research interests include clustering, feature selection, machine learning and its application fields.

Tetsuya Sakurai is a professor of the Department of Computer Science, and the director of Center for Artificial Intelligence Research (C-AIR) at the University of Tsukuba. He is also a visiting professor at the Open University of Japan, and a visiting researcher of Advanced Institute of Computational Science at RIKEN. He received a Ph.D. in computer engineering from Nagoya University in 1992. His research interests include high performance algorithms for large-scale simulations, data and image analysis, and deep neural network computations. He is a member of the Japan Society for Industrial and Applied Mathematics (JSIAM), the Mathematical Society of Japan (MSJ), Information Processing Society of Japan (IPSJ), Society for Industrial and Applied Mathematics (SIAM).