



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Data Quality Improvements for Multiclass Classification

Juhui Lee

Department of Management Engineering

Graduate School of UNIST


Data Quality Improvements for Multiclass Classification

A thesis/dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Juhui Lee

07/03/2020

Approved by



Advisor

Sungil Kim

Data Quality Improvements for Multiclass Classification

Juhui Lee

This certifies that the thesis/dissertation of Juhui Lee is approved.

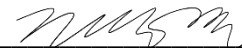
07/03/2020

signature



Advisor: Sungil Kim

signature



Sungil Kim: Thesis Committee Member #1

signature



Chiehyeon Lim: Thesis Committee Member #2

signature



Junghye Lee: Thesis Committee Member #3

Abstract

The success of machine learning (ML) is based on data quality and representation. Noisy and unreliable data with much irrelevant and redundant information makes the learning phase of ML even more difficult. A feature transformation is well known as an essential step to overcome data quality problems in ML problems.

A proper feature transformation depending on the application domain would provide more understandable information. Creating new meaningful values contributes to a better understanding of a classifier. The present paper focuses on two data quality problems are representative issues of multiclass classification problems that arise when collecting or experimenting with data; a class imbalance of ordinal data, and drift compensation of Electronic nose (E-nose). First, we proposed truncated Singular Value Decomposition for Multiclass Classification (SVDMC) for imbalanced ordinal classification. The proposed model is novel in that it can handle the issue without modifying the class distribution. Next, we propose a novel compensation method to address sensor drift under batch experiments. The proposed model is structured based on the nonlinear parametric function of experimental factors.

To improve data quality for multiclass classification, we applied Genetic Algorithms (GAs) to optimize both techniques to improve the classification performance. The results based on simulations and real datasets show that the classification performance is significantly improved after the GAs optimized feature transformations than when using classification models alone.

Key words: Data Preprocessing, Feature Transformation, Genetic Algorithm, Class Imbalance, Gas Sensor Classification, Singular Value Decomposition, Sensor Drift, Correction Model, Data Denoising

Table of Contents

Table of Contents	6
List of Figures	8
List of Tables	9
1. Introduction.....	1
1.1 Background: Data Quality Impact on Multiclass Classification.....	1
1.2 Purpose and Outline	2
2. Literature Reviews	4
2.1 GA Optimized Feature Transformation.....	4
3. SVD Truncation for Multiclass Classification (SVDMC).....	6
3.1 Introduction	6
3.2 Literature Reviews	7
3.3 Methodology	8
3.3.1 Additive Perturbation Model	9
3.3.2 Singular Value Decomposition (SVD).....	9
3.3.3 Denoising by Truncating Singular Values.....	10
3.3.4 Truncated SVD using GAs for Multiclass Classification	11
3.4 Description of Experimental Setup	12
3.5 Simulation Study	15
3.6 Real Data Example.....	19
4. Sensor Drift Compensation for Mixed Gas Classification under Batch Experiments	21
4.1 Introduction	21
4.2 Literature Reviews	21
4.3 Methodology	22
4.3.1 Sensor Drift Compensation Model	23
4.3.2 Class Separability Criterion of GAs.....	23

4.4	Description of Experimental Setup	25
4.5	Simulation Study	26
4.6	Real Data Example.....	29
5.	Conclusion	34
5.1	Summary and Contributions.....	34
5.2	Limitations and future research.....	34
	References.....	36

List of Figures

Figure 1: Comparison of wrapper and filter models	4
Figure 2: Comparison of SVD truncation approaches	11
Figure 3: Pseudocode of SVDMC	12
Figure 4: Confusion matrix in binary classification problem	13
Figure 5: Distributions of the disaster data	19
Figure 6: Pseudocode of the proposed drift compensation algorithm.....	25
Figure 7: Block diagram of the simulation data preprocessing.....	26
Figure 8: The distributions of simulation data1	27
Figure 9: The distributions of simulation data2	28
Figure 10: A measured gas with reference voltages and effective voltages.....	30
Figure 11: A scatter plot showing the steady state values of sensor 1 by the group.	31
Figure 12: Scatter plots of original sensor 1 values and corrected sensor 1 values	32
Figure 13: Scatter plots of original sensor 8 values and corrected sensor 8 values	32

List of Tables

Table 1: Comparison of empirical classification results with $\sigma = 1.0$	16
Table 2: Comparison of empirical classification results with $\sigma = 1.5$	16
Table 3: Comparison of empirical classification results with $\sigma = 2$	16
Table 4: Danger values and information of β_1 and β_2 of the first simulation.	16
Table 5: Comparison of empirical classification results with $r = 3$	18
Table 6: Comparison of empirical classification results with $r = 5$	18
Table 7: Danger values and information of β_1 and β_2 of the second simulation.....	18
Table 8: Comparison of empirical classification results of synthetic multiclass data.....	18
Table 9: Experimental classification results of disaster data on $wkNN$ ($k=7$) algorithm ...	19
Table 9: Comparison among true parameters α_{1j} and the estimated ones $\hat{\alpha}_{1j}$ for $j = 1, \dots, 5$	28
Table 10: Result of classifications. The average of classification performances on ANN set was 0.41.	29
Table 12: Summary of classification performance.....	31

1. Introduction

1.1 Background: Data Quality Impact on Multiclass Classification

The most important factor in Machine Learning (ML) is data quality. Machine learning algorithms automatically extract knowledge from data. However, their success in knowledge extraction depends on not only their learning performance skills, but also the quality of the data. If the data contains a lot of irrelevant information, algorithms may not produce accurate and comprehensible results, or they may not learn the data at all. Therefore, in ML, improving the data quality through data preprocessing is first and foremost. Preprocessed data provide better discrimination as shown by S. Kotsiantis [1].

Preprocessing techniques can be divided into seven categories: instance selection and outlier detection, processing missing values, discretization, data normalization, feature selection, and feature transformation.

- Instance selection and outlier detection: techniques detecting some possible data quality problems and removing instances with excessively deviating instances.
- Processing missing values: techniques for completing unknown feature values. It is one of the most critical issues since most data in the real world is incomplete and contains missing values.
- Discretization: Preprocessing to convert real values to integers. The integer feature values sometimes make algorithm learning more effective and faster.
- Data normalization: a "scaling down" transformation of the features. Normalization adjusts the values to lower values or changes the values to a common scale without distorting differences in the ranges of the values.
- Feature selection: the process of identifying meaningful features and reducing dimension by removing irrelevant features. This process makes the algorithms run faster and more effectively.
- Feature transformation: techniques for constructing new features from the original feature values, resulting in improvement in the representation of the data.

In supervised learning, the impact of feature transformation is very significant. Transformed features can produce more concise and accurate information. Creating new meaningful values

contributes to better understanding of a classifier. Therefore, appropriate feature transformation related to the application domain can provide a high classification accuracy.

This dissertation comprises two data quality issues related to multiclass classification problems. They are representative issues that may arise when collecting or experimenting with data; Class imbalance, and drift compensation. First of all, class imbalances, which have been studied by a lot of papers, can occur during data collection. Class imbalance occurs during data collection. Suppose, for example, that we are detecting a process defect. Usually, the number of data in the abnormal group collected is very small. There exist difficulties in collecting data in the abnormal group, which would generate imbalanced information among different groups. Classifiers have low classification accuracy for rare events due to the imbalanced information. This example shows the effect of class imbalance.

On the other hand, if we generate data through experiments, we have some control over the effects of class imbalance by planning a balanced class distribution in the experimental setup. Suppose, for example, that we are experimenting with gas detection with an Electronic nose (E-nose). In the experimental setup, we can plan to conduct the same number of experiments for different types of gases, which would prevent happening the class imbalance. However, during experimentation, various environmental factors can affect data properties. For example, fluctuations in temperatures and humidity conditions cause an unexpected chemical reaction inside sensors, and sensor values have patterns independent of the gas property. This example corresponds to sensor drift that is most well-known problem in E-nose. It makes unexpected sensor values, which makes a classifier recognize the odor type of sensor difficult. It impairs the data quality and reliability of sensor array data. This dissertation proposed two feature transformation techniques to deal with the two data quality problems of data collection and production on multiclass classification problems.

1.2 Purpose and Outline

The objective of this dissertation is to improve multiclass classification performance by suggesting feature transformation techniques optimized by Genetic Algorithms (GAs). Feature transformations are sometimes complicated and contain parameters to be determined. Heuristic methods have been mainly applied for determining the parameters. Especially, GAs, one of the heuristic methods, successfully search optimal or local optimal values for feature transformation as in [2, 3, 4]. GAs belong to a class of probabilistic algorithms based on natural selection and natural genetics to perform parallel searches in complex search spaces [5]. It reflects the natural selection process of selecting the best-fitted object by reproduction to create the offspring of the next generation. The fit of the object is calculated from the given fitness function corresponding to the optimization

problem. There are several advantages to it. It does not require derivative information, is faster and more efficient than traditional methods, optimizes continuous and discrete functions and multi-objective problems, and is useful for large search spaces with many parameters. In pattern classification, GAs have been used for parameter tuning [6], feature transformation [2, 4], and feature selection [7, 8, 9, 10]. Based on the properties of GAs, we proposed two genetic algorithm optimized feature transformation techniques for supervised learning on two specific domains; imbalanced ordinal classification, gas identification with sensor drift.

First, we propose Singular Value Decomposing truncation for Multiclass Classification (SVDMC) to deal with imbalanced ordinal data. The ordinal data refers to data that contains a meaningful order between classes, but the distance between categories cannot be strictly quantified. Also, they often suffer from noisy ratings. Not all input ratings are reliable, and noisy ratings can be detrimental to the quality of the trained model. Therefore, the ordinal data classification problem is quite tricky. Considering the two ordinal data properties, the paper addresses the class imbalance by reducing noise and class overlap instead of applying a sampling technique. To do this, we develop SVD truncation that is suitable for noise reduction and combine it with GA also helps the model find the ultimate data space. The main contribution of this model is a new pre-processing method for classifying imbalanced ordinal data using single value singular value reduction.

Second, we deal with sensor drift, the most well-known problem in the classification of electronic nose data. Sensor drift generates unexpected sensor values triggered by several factors such as aging, ambient temperature, humidity, pressure, and poisoning. It impairs the data quality and reliability of sensor array data. Therefore, drift compensation can improve system reliability and increase the accuracy of gas recognition through the sensor array system. The paper proposes a model that reconstructs drifted values based on the nonlinear parametric function. If there is not enough gas data for gas analysis, the regression algorithm cannot find a solution, so we apply GAs to find the parameters. It is an innovative preprocessing technique for sensor drift compensation.

The rest of the dissertation is organized as follows. Chapter 2 presents the research background on GA optimized feature transformations. Chapter 3 and Chapter 4 cover the two issues that make up the dissertation work. Each chapter includes the introduction, literature reviews, methodology, experiments, and results. Finally, Chapter 5 concludes with the contributions and limitations of the dissertation and suggests future research directions.

2. Literature Reviews

2.1 GA Optimized Feature Transformation

Genetic Algorithms (GAs) have been applied as a method of determining parameters. GAs are probabilistic algorithms based on natural selection and natural genetics to perform parallel searches in complex search spaces [5]. It reflects the natural selection process of selecting the fittest chromosome (or genetic property) by reproduction to create the offspring of the next generation. In GAs, the fitness of a chromosome is calculated from a given fitness function. Their most important application is machine learning, successfully applied in parameter tuning [6], feature transformation [2, 3, 4], and feature selection [9, 7, 10, 8].

GAs have several properties such as simple programming capabilities, flexibility in organizing fitness function, and robustness to the input data [11]. These features are effective in cases where it is difficult to solve or even express the objective function for parameter estimation. The objective function for SVDMC is too complicated to express explicitly or implicitly because it is associated with singular values in the class matrix and the objective function of a classifier. So, we need an optimization solver that provides a simple programming capability and is flexible in configuring fitness functions. Meanwhile, the objective function for sensor drift compensation can be expressed explicitly. However, drifted sensor data often do not facilitate finding such a solution. In this case, traditional hill climbers have not been able to find an optimal solution. GAs, on the other hand, have been succeeded in finding robust solutions, which is achieved by operators like selection, crossover, and mutation. Based on the properties of GAs, we proposed two genetic algorithm optimized feature transformation techniques for supervised learning; imbalanced ordinal multiclass classification, sensor drift compensation for mixed gas classification under batch experiments. GA optimized feature transformation techniques can be categorized into two groups: wrapper models

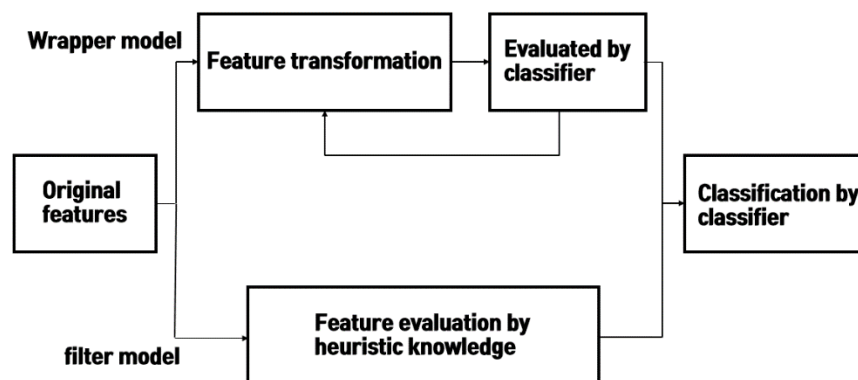


Figure 1: Comparison of wrapper and filter models [4]

and filter models. Wrapper models transform features by feedback from the classifier evaluation, while filter models do this in a heuristic way. Figure 1 shows the difference between these two approaches. An example of wrapper models, Prakash and Murty [2] suggested the combination of GAs with Principal Components Analysis (PCA). GAs select an optimal subset of PCs to get the best performance. In the paper [4], the authors proposed a new feature weighting and selection technique based on wrapper model. GAs determine the weights of the features from 0.0 to 10.0. When the determined weight is 0.0, the technique completes the feature selection. Raymer *et al.* [12] presented a feature extraction method in which GAs weight according to the importance of the feature. A gaussian windowing function was proposed to improve the classification accuracy of an odor classification in [13]. The authors developed the model using GAs to determine the optimal parameters for feature extraction. It also has a positive effect on time drift. Compared to the number of studies on wrapper models, few studies were done on the filter approach using GAs. The paper [3] proposed new framework of GAs that optimize the S-transform of perturbations in electrical signals. The fitness function of GAs is to maximize the energy concentration.

The proposed SVDMC adjusts the data space of classes differently to reduce the imbalance effect, which corresponds to wrapper models. GAs determine the amount of data space to reduce to maximize classification performance. The feature transformation for drift compensation, on the other hand, corresponds to filter models. GAs search for a solution that optimizes a specific fitness function, not a classification performance. We suggest a fitness function that is effective on sensor drift compensation. It measures class separability based on the ratio for the trace of the between-class scatter matrix (S_B) and that of the within-class scatter matrix (S_W). This class separability function will be explained in more detail in Section 4.3.2.

To the best of our knowledge, there are little GA optimized feature transformation techniques for class imbalance and sensor drift. So, we also review the preprocessing methods for the ordinal classification of imbalanced data and the gas identification with sensor drift in the following sections. As a comparison model, we select one GA optimized feature transformation technique most relevant to each proposed model, and well-known preprocessing techniques developed in two areas.

3. SVD Truncation for Multiclass Classification (SVDMC)

3.1 Introduction

There are a lot of ordinal classification problems in real life such as risk assessment, sentiment analysis, diagnostic classification, and image analysis in [14] and [15]. The multiclass ordinal data inherently have low probability classes (e.g., the most dangerous disaster class or diagnosed class with cancer). In such cases, it is important to detect unusual cases, but it is challenging because they have very few observations compared to other groups [16]. Although much attention has been given to general imbalanced data, its ordinal counterpart has not received the same attention. This paper analyzes ordinal data with class imbalance problems.

There are mainly three approaches of solutions: the external approach of preprocessing data [17, 18, 19, 20, 21], the internal approach of modifying algorithms [15, 22, 23], and cost-sensitive learning [24, 25]. Cost-sensitive learning helps to accurately classify minority classes by assigning higher misclassification costs to minority class than majority class. However, it takes a lot of time to find an appropriate cost value for each class, and misclassification costs are often unavailable [26, 27]. The internal approach aims at modifying or modeling new algorithms to classify imbalanced data, but they cannot deal with the issues at the most basic level and their performance depends on the data. Of the three approaches, the external approach is most suitable for reducing the impact of class imbalance [28]. Its main advantage is its versatility, as it does not depend on a specific classifier. Besides, it can also be flexibly integrated into other types of approaches. The external approach mainly consists of sampling techniques, undersampling or oversampling techniques. Undersampling techniques aim at deleting some majority class samples and oversampling techniques generating new minority class samples. However, they do not seem to take into the properties of an ordinal imbalanced data account.

There are two considerations for ordinal data [29]. First, not all input ratings are reliable, and noisy ratings are detrimental to the quality of the trained model. Second, the label ordering of the data is meaningful in the topology of the sample space. Many studies for ordinal-imbalanced data consider only the second feature, but the first feature is significant in data preprocessing. Sampling techniques are based on data class reliability. Unless all input ratings are reliable, oversampled data will have more noisy ratings, and undersampled data will suffer from a lot of loss information. Therefore, we propose a new external approach not modifying a skewed class distribution. Instead, we aim to improve data quality and the reliability of the ratings by reducing noise in all input data to handle

ordinal-imbalanced data.

Our goal is to improve data quality by reducing data noise and overlap regions to handle imbalanced ordinal data. Many researchers have asserted that the problem of the imbalance is not by itself but other data quality problems while pointing out the importance of reducing noise and overlap areas in imbalanced data classification [30, 31]. Data noise and overlap area harm the learning of a classifier on imbalanced ordinal data [30]. The authors [31] states that if samples overlap in skewed data, it can be challenging to train a classifier effectively. We address the noise and overlap area reduction of imbalanced ordinal data by singular value shrinkage.

Data noise reduction through Singular Value Decomposition (SVD) has been successful in [32, 33, 34, 35]. SVD is suitable for noise reduction because it allows one to understand the characteristic of the data matrix. Matrix factorization allows one to quantify the relationships within the samples, within the attributes, and between samples and attributes with singular values. By applying information on singular values, it is possible to reduce noise and overlap area in data space. The proposed model takes the class order into account and does not reverse the existing order of the transformed input data. The main contribution is that the proposed novel preprocessing method does not modify a class distribution. Instead, it reduces the additive noise of the data and overlap area using singular value shrinkage. To the best of our knowledge, this is the first attempt at utilizing SVD to solve imbalances between ordinal classes.

3.2 Literature Reviews

In related works, the existing methods for resolving class imbalance can be broadly categorized into three approaches: the external approach of preprocessing training sets to balance class distribution [17, 18, 19, 20, 21], the internal approach of creating or modifying algorithms to accommodate class imbalance problems [15, 22, 23], and a combination of the two [24]. Among the three approaches, data preprocessing techniques have been widely used and recognized as appropriate for reducing the impact of class imbalances [28].

The most known approaches of data preprocessing are undersampling and oversampling techniques that regenerate a training dataset to create a balance between classes by eliminating some examples from data or creating synthetic data. The most well-known undersampling technique is the Wilson's Edited Nearest Neighbor (ENN) rule [36], which edits preclassified samples and reduces the number of samples by deleting samples, wherein at least two of the three nearest neighbors are in different classes. The most well-known oversampling technique is the Synthetic Minority Oversampling Technique (SMOTE) [37]. SMOTE synthesizes samples from minority classes

based on the distance of the nearest samples. Many adaptive oversampling algorithms are modified versions of the SMOTE [38, 39, 40, 41, 16, 42, 43]. Among the adaptive algorithms is SMOTE and Cluster-based Undersampling Technique (SCUT) [43], which is a combination of SMOTE and cluster-based undersampling. However, these algorithms are designed for imbalanced general data, not imbalanced ordinal data.

Recently, some preprocessing methods have been proposed in an ordinal context. Domingue *et al.* [17] proposed four oversampling techniques based on ordinal data; 2 to 3, feature by feature, centroid-based, and principal component analysis (PCA)-based. They generate some samples from minority classes based on the main characteristics such as the mean, median, and principal components (PCs) of class data. The same authors proposed the iterative oversampling techniques called InCuBAte that outperformed the four methods [18]. The authors argued that the performance of sampling improves as the function selection proceeds. The main algorithm is as follows. First, ReliefF performs feature selection on training data. The dataset is then duplicated, and a classifier learns from the duplicated dataset. Sample are then generated and added to the training set, where the class of the samples are assigned labels predicted by the classifier. These steps are iteratively repeated until enough samples have been generated. Authors in [19] proposed an ordinal oversampling method from a graph-based perspective. They demonstrated a good synergy with support vector ordinal regression. Nekooimehr and Lai-Yuen [21] presented adaptive semi-supervised weighted oversampling (A-SUWO) for imbalanced binary classification. A-SUWO oversamples each sub-cluster using a semi-supervised hierarchical clustering approach. The main algorithm of A-SUWO also contains a denoising step. Cluster-based weighted oversampling for ordinal regression (CWOS-Ord) [20] is a modified version of the A-SUWO developed by the same authors. Most preprocessing techniques for the ordinal classification or regression favor the oversampling approach with denoised data, but few consider reducing overlapping regions and noise. Therefore, our approach is a new paradigm for dealing with imbalanced ordinal classification.

3.3 Methodology

We propose an SVD truncation method for imbalanced ordinal data classification. In Section 3.3.1 and Section 3.3.2, we briefly review the additive perturbation model and SVD. In Section 3.3.3, we discuss a method for determining a proper threshold for successful SVD truncation. In Section 3.3.4, we present our SVD truncation algorithm for multiclass classification (SVDMC) optimized by GAs.

3.3.1 Additive Perturbation Model

Analysis of the additive perturbation model relates to the spectrum changes after a small perturbation to a matrix. It is one of the main concepts in the analysis of SVD and spectral methods. In other words, it is important to understand how different the data spaces of X and $X + Z$ are, where Z is a perturbation model. The additive perturbation model is applicable to a wide range of problems, including matrix denoising, clustering, community detection in bipartite, singular space estimation of matrix completion, and canonical correlation analysis as demonstrated in [44, 45, 46]. In our paper, we analyze ordinal data noise with the additive model.

Let Y denote an $n \times m$ data matrix. The additive perturbation model decomposes the data matrix into two parts [47]:

$$Y = X + \sigma Z \quad (1)$$

where X denotes the noise free data and Z denotes noise. The noise is assumed to be independent identically distributed zero-mean Gaussian with unit variance while σ is a positive real value. For matrix decomposition, SVD can be easily applied in the additive perturbation model to extract the noise free data X from the data matrix Y .

3.3.2 Singular Value Decomposition (SVD)

Formally, the singular value decomposition of an $n \times m$ matrix Y is into the product of three matrices of the form $U\Sigma V^T$. U is an orthogonal $n \times n$ matrix having columns $\mathbf{u}_i = [u_{1i} u_{2i} \cdots u_{ni}]^T$ for $i = 1, \dots, n$ and V is an $m \times m$ matrix with orthonormal vectors, $\mathbf{v}_j = [v_{1j} v_{2j} \cdots v_{mj}]^T$ for $j = 1, \dots, m$. Σ is an $n \times m$ rectangular diagonal matrix and has non-negative diagonal elements that are singular values y_i of Y . The singular values are ordered so that $y_i \geq y_j$ for all $i < j$. In the case of the rank of r , Y can be shown as a combination of r matrices [47]:

$$Y = U\Sigma V^T = \sum_{i=1}^r y_i \mathbf{u}_i \mathbf{v}_i^T. \quad (2)$$

The matrices U and V contain information about data observations and data attributes respectively. The similarity of columns within U or V indicates that the corresponding observations or attributes are closely related. Also, the singular values y_i of the matrix Σ indicate

the importance of the corresponding orthonormal vectors \mathbf{u}_i and \mathbf{v}_i^T , called basis vectors. The singular values reflect the variance of the data captured by the corresponding basis vectors. The basis vectors are arranged in order of the magnitude of singular values. The first basis vector with the largest singular value is in the direction of the largest data variance. The second basis captures the orthogonal direction with the second greatest variance, and so on. In geometrically, the singular values of the matrix Y indicate the lengths of the semi-axes of an ellipse so that the truncated matrix with smaller singular values has the smaller ellipse. For a finite $n \times m$ matrix Y , the semi-axes of an m -dimensional ellipsoid in an n -dimensional space are represented as singular values of the matrix Y . Then, reducing the lengths of the semi-axes makes the elements of matrix Y closer. For multiclass classification problems, observations of the same class are relocated nearby through SVD truncation with a set of reduced variances.

3.3.3 Denoising by Truncating Singular Values

A denoised matrix is estimated using SVD truncation by removing some meaningless bases. Bases represent observations as linear combinations composing a matrix, and the rank of a matrix is equal to the number of its bases. The matrix can have the optimal rank by eliminating the meaningless bases. When the $n \times m$ matrix Y of rank r has singular values y_i for $i = 1, \dots, r$. The optimal rank of p can be estimated using a threshold δ , where

$$y_1 \geq y_2 \geq \dots \geq y_p \geq \delta \geq y_{p+1} \geq \dots \geq y_r.$$

Then, the singular values y_i are truncated for $i > p$ and the truncated SVD of the matrix Y becomes the matrix X ;

$$X = \sum_{i=1}^p y_i \mathbf{u}_i \mathbf{v}_i^T.$$

It is called truncated SVD using a hard threshold. Some meaningless bases are deleted with the threshold δ , but some noise included in the data space related to some remaining singular values cannot be reduced. If noise occurs in every vector space composed of singular vectors, it is recommended to use a soft threshold for truncating SVD. To delete more noise from every vector space composed of singular vectors, a soft threshold is adopted in our experiments. Recall that for $y_i > 0$,

$$X = \sum_{i=1}^r \max(0, y_i - \beta) \mathbf{u}_i \mathbf{v}_i^T. \quad (3)$$

We truncate some irrelevant bases and shrink the data space using the soft threshold in Eq.(3). This approach has been widely used for scalar and vector noise removal because of its simplicity and several optimality properties [48, 49, 50]. There are many studies estimating the threshold β of the matrix X but most of them do not consider the shape of the data matrix Y . For classification, it is essential to consider the property of the data space. Thus, it is difficult for them to provide the proper threshold for classification. Therefore, we aim to obtain the optimal with the results obtained by learning the input data. In the following subsection, we introduce a new truncating singular values method for multiclass classification called the SVDMC, and a learning process for finding the optimal threshold.

3.3.4 Truncated SVD using GAs for Multiclass Classification

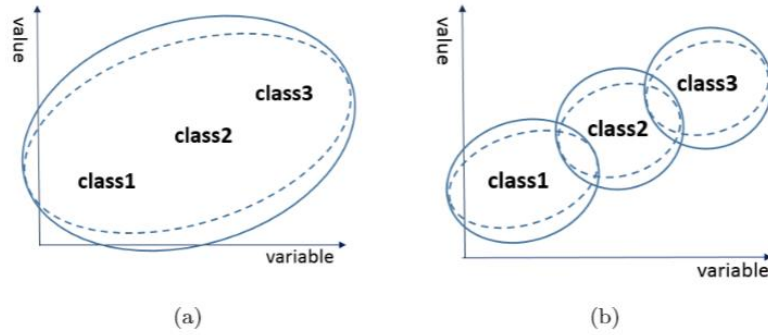


Figure 2: Comparison of SVD truncation approaches: (a) existing approach [60] and (b) SVDMC. The solid line represents the untruncated data space, and the dotted line represents the truncated data space.

The proposed SVD truncation method, SVDMC, applies SVD truncation to every class, not just to the entire data matrix. This is because that the standard SVD truncation to the data would increase overlap regions. Figure 2 illustrates the difference between SVD truncation on entire datasets and SVD truncation on every class separately. SVDMC prevents overlapping regions from becoming larger by denoising every class. An important question then arises as to how the soft threshold can be applied in SVDMC.

Suppose we have a total number of K ordinal classes. Let $Y^{(k)}$ denote an $n_k \times m$ data matrix consisting of data points belonging to the k th class for $k = 1, \dots, K$. According to Eq.(1), the additive perturbation model decomposes data matrices in the following manner.

$$Y^{(k)} = X^{(k)} + \sigma^{(k)}Z,$$

where $X^{(k)}$ and $\sigma^{(k)}Z$ denote the noise-free and the noise matrix of k th class respectively. For each $Y^{(k)}$, the corresponding value of β_k determines how shrunk the data space becomes. We construct

K thresholds having parameters β_k , $k = 1, \dots, K$ for truncating the K classes.

Individual truncated data spaces of classes have an impact on the classification performance, and so does the combination of truncated spaces among classes. That is, β_k are not independent of each other. Thus, our goal is to find the optimal combination of β_k , that is, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)^T$ that maximizes the classification performance on the imbalanced data. The proper threshold $\boldsymbol{\beta}$ helps classifiers determine unbiased decision boundaries that can provide good classification performance in every class. Therefore, we set the optimal $\boldsymbol{\beta}$ as the value maximizing one of the performance measures on imbalanced data. To do find the optimal parameter, we apply GAs as we have been discussed. Figure 3 illustrates the pseudocode of SVDMC optimized by GAs.

Algorithm 1: SVDMC's algorithm

```

1: procedure SVDMC( $D$ ) ▷ Input is a training set
2:   procedure BETA( $\beta = (\beta_1, \dots, \beta_K)$ )
3:     sum,  $r = 0$  while  $r < 5$  do
4:       Arbitrarily choose subtraining & subtest from  $D$  where  $|\text{subtraining}| = 0.9 * |D|$  and
        $|\text{subtest}| = 0.1 * |D|$ 
5:       sub $Y^{(k)}$  is the data matrix of class  $k$  for  $k = 1, \dots, K$ .
6:       sub $X^{(k)} \leftarrow \sum_{i=1}^{r^{(k)}} \max\{0, y_i^{(k)} - \beta_k\} \mathbf{u}_i^{(k)} \mathbf{v}_i^{T^{(k)}}$  for  $k = 1, \dots, K$ .
7:       sub $X$  is the data matrix containing sub $X^{(1)}, \dots, \text{sub}X^{(K)}$ .
8:       results  $\leftarrow$  classifier(trainset = sub $X$ , valid = subtest)
9:       sum = sum + results
10:       $r = r + 1$ 
11:
12:     return average of sum
13:   end procedure
14:    $\hat{\beta} = \text{GA}(\text{BETA})$  ▷ Find  $\hat{\beta}$  maximizing the value of the function BETA
15:    $Y^{(k)}$  is the data matrix of class  $k$  for  $k = 1, \dots, K$ .
16:    $X^{(k)} \leftarrow \sum_{i=1}^{r^{(k)}} \max\{0, y_i^{(k)} - \hat{\beta}_k\} \mathbf{u}_i^{(k)} \mathbf{v}_i^{T^{(k)}}$  for  $k = 1, \dots, K$ .
17:    $X$  is the data matrix containing  $X^{(1)}, \dots, X^{(K)}$ .
18:   return  $X$ 
19: end procedure

```

Figure 3: Pseudocode of SVDMC

3.4 Description of Experimental Setup

This section describes the experimental setup such as comparison techniques, classification algorithms, and classification performance measurements.

True Class \ Prediction	Positive Prediction	Negative Prediction
	Positive Class	True Positive (TP)
Negative Class	False Positive (FP)	True Negative (TN)

Figure 4: Confusion matrix in binary classification problem

The performance of SVDMC is evaluated by comparing four existing preprocessing methods for addressing the class imbalance and one singular values-based method using GAs. For preprocessing methods for resolving class imbalance, two oversampling algorithms (SMOTE [37] and InCuBAte [18]), one undersampling algorithm (ENN [36]), and one hybrid algorithm (SCUT [43]) are considered. For the singular values-based method, the wrapper method of GA optimized feature transformation [2] is considered, which is referred to as GAPCs in this paper. Note that over/under/hybrid sampling algorithms adjust the number of data points in each class but truncating singular values-based methods maintain the original distributions of data points over classes.

Consequently, we have seven types of training sets: the original training set, SVDMC set, GAPCs set, SMOTE set, InCuBAte set, SCUT set, and ENN set. Preprocessed datasets are learned and evaluated by the ordinal classification algorithm $wkNN$ for the ordinal data version [51]. The $wkNN$ is implemented with Euclidean distance and the triangular kernel for weighting at $k = 7$.

For reference, we applied a genetic algorithm as follows. The crossover rate is 80%, and the mutation rate is 0.1. GAs run 100 generations with a population of 100. Evolution stops when there has been no improvement over the past 30 generations. GAs were implemented in ‘GAparsimony [52]’ that is the R package. All tests have been run on R studio cloud with R 3.5.3 version, which is the cloud provided R studio. The computation time was calculated in Section 3.6.

In this paper, we applied G-mean, and MAUC as performance measures. The geometric mean (G-mean) and the Area Under the receiver operating characteristic Curve (AUC) are commonly used to measure the performance of multiclass imbalance problems [53]. Before explaining the measures, we will review the confusion matrix.

Let’s consider a binary classification problem where data belong to either Class 1 or Class 2. When the class imbalance occurs, the majority class is also called the negative class. The other one is called the positive class. In our case, we assign Class 1 as the negative class and Class 2 as the

positive class. The confusion matrix provides insight into the distribution of correct/incorrect data instances in a class, and what types of errors occur, as well as the performance of the predictive model. Figure 4 illustrates a summary of the confusion matrix. It is the framework that formulates G-mean, and AUC.

The true positive rate and the false positive rate are essential concepts in the confusion matrix. The true positive rate (the false positive) is the value calculated as the total number of true positive (false positive) predictions divided by the sum of the true positives (false positives) and the false negatives (true negatives). The true positive rate is called sensitivity or recall. The value of $1 -$ the false positive rate is referred to as the specificity. In summary, they can be formulated as follows;

$$\text{True positive rate} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

$$\text{False positive rate} = \frac{\text{FP}}{(\text{FP}+\text{TN})}$$

$$\text{Specificity} = 1 - \text{false positive rate}$$

The G-mean is then the geometric mean of the sensitivity and the specificity. AUC is the area under the ROC curve that is a plot with x-axis as the false positive rate and y-axis as the true positive rate. The domains of both measures are from 0 to 1, and the maximum denotes the best classification performance. In general, the minimum value of AUC is 0.5.

The measures are originally designed for binary classification problems so that we use the extended G-mean [54] and extended AUC, called MAUC [55], for multiclass problems. The extended G-mean is defined as the geometric mean of the true positive rates of all classes, and the extended AUC is defined as the average AUC of all pairs of classes. We determine the parameters of SVDMC with values that maximize G-mean result. This is because G-mean is more sensitive to changes in the confusion matrix than MAUC. It is helpful for GAs to find an optimal solution.

In addition to these performance measures, we measure the overlap regions using a distance-based metric called Danger, which is defined as follows.

$$\text{Danger}(y, \mathbf{X}) = \frac{1}{q} \sum_{i=1}^n \sum_{l=1}^q |y_i - y_{i(l)}|,$$

where $y_i \in \{1, 2, \dots, K\}$ is an ordinal class label and $y_{\{i(l)\}}$ represents the class label of the l -th nearest neighbor in the feature space defined by \mathbf{X} . This metric measures the homogeneity of the data set. It is designed to increase the metric value as more data points are in the overlapping regions between classes. If data points are located close to data points having the same class label, the value of the metric is close to zero. Danger is used to analyze the characteristic of the original data as well as to quantify the effect of SVDMC. Note that a low Danger value does not guarantee good classification performance. Rather, if the Danger value of the transformed data is very different from the Danger value of this data, the transformed data is likely to have lost the characteristics of the data. The purpose of introducing Danger measure is to quantify the overlap area characteristics of this data and to analyze the effect of the proposed model based on this.

For evaluation, data are divided into a training set (80 percent of the data) and a test set (20 percent of the data) while preserving the degree of imbalance in the class distribution. For the wrapper models, five validation sets are randomly selected as 10 percent of the training set for tuning the hyperparameter. Data is randomly dichotomized 50 times into a training set and a test set to avoid biased outcomes. Then, empirical results are obtained by averaging the results.

3.5 Simulation Study

To study the impact of SVDMC on low-rank ordinal data, we conducted three simulation studies: a) binary ordinal classification with different σ , b) binary ordinal classification with different rank r , and c) multiclass ordinal classification. For the first simulation, we expect that the higher the amount of noise in the data, the better the classification performance improvement of the proposed model. We expect that the proposed model performance will perform well on low rank as well as other data for the second simulation. For last, we confirm the performance of the proposed model in multiclass classification.

When generating synthetic data, for simplicity, we assume that $Y^{(k)} = X^{(k)} + \sigma^{(k)}Z$ and the rank of the matrix $X^{(k)}$ is r for all k . The entries of the matrix $X^{(k)}$ are randomly generated from the uniform distribution with an unknown interval $[a_k, b_k]$, where $a_k < b_k$ and $b_k < a_{k+1}$. In addition, the entries of the matrix Z are randomly generated from a Gaussian distribution with zero mean and unit variance.

Table 1: Comparison of empirical classification results of synthetic binary data with $\sigma = 1.0$

<i>wkNN</i>	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.50	0.55	0.50	0.52	0.50	0.50	0.55
	± 0.01	± 0.06	± 0.02	± 0.04	± 0.03	± 0.05	± 0.01
G-mean	0.08	0.52	0.18	0.44	0.17	0.53	0.00
	± 0.10	± 0.07	± 0.14	± 0.08	± 0.13	± 0.07	± 0.00

Table 2: Comparison of empirical classification results of synthetic binary data with $\sigma = 1.5$

<i>wkNN</i>	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.49	0.54	0.50	0.50	0.50	0.50	0.51
	± 0.01	± 0.04	± 0.03	± 0.05	± 0.03	± 0.06	± 0.00
G-mean	0.04	0.51	0.15	0.44	0.21	0.50	0.00
	± 0.09	± 0.06	± 0.18	± 0.09	± 0.12	± 0.07	± 0.00

Table 3: Comparison of empirical classification results of synthetic binary data with $\sigma = 2$

<i>wkNN</i>	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.49	0.52	0.50	0.50	0.50	0.50	0.50
	± 0.01	± 0.05	± 0.03	± 0.05	± 0.02	± 0.06	± 0.00
G-mean	0.04	0.49	0.09	0.41	0.19	0.47	0.00
	± 0.09	± 0.07	± 0.13	± 0.07	± 0.12	± 0.07	± 0.00

Table 4: Danger values and information of β_1 and β_2 of the first simulation.

	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 2$
Danger(original)	0.13	0.14	0.14
Danger(SVDMC)	0.06	0.07	0.07
β_1	20.73 ± 1.20	30.94 ± 1.74	41.67 ± 2.37
β_2	5.41 ± 1.19	7.04 ± 2.27	9.86 ± 2.68

a) *Binary ordinal classification with different σ* : We conducted binary ordinal classification for $\sigma = 1.0, 1.5, 2.0$. Denote the majority class as Class 1 and the minority class as Class 2. Class 1 has 1000 observations with $r = 1$ and $m = 10$. Class 2 has 100 observations with the same r and m . A low rank matrix has a ratio $\frac{r}{m} \leq 0.1$ in general, which enables us to choose appropriate r and m values. Tables 1, 2, and 3 display the results for $\sigma = 1.0, 1.5, 2.0$ respectively.

Our goal is to reduce noise in the overlap area. The more noise there is, the more effective our model will be. The rate of increase in the classification performance of SVDMC according to the rate of the best increase in the classification performance of the compared models were 100%, 250%, and 300% (see Table 1, Table 2, and Table 3), respectively. The rates with respect to G-mean were about 98%, 102%, and 105% (see Table 1, Table 2, and Table 3), respectively. Therefore, the proposed model contributed more to improving classification performance for noisy data than other comparative models. The results confirm that the proposed model plays a proper role. GAPCs, SMOTE, and InCuBAte model show slightly improved performance results. SCUT, a hybrid technique, shows superior results over the two oversampling models. The undersampling technique, ENN, no longer seems to fit the imbalance issue.

In our simulation data, the higher the sigma, the higher the Danger value in Table 4. Original data space is more deformed by the additive noise and more noise accumulated in the overlap area. In this case, the proposed model will enhance the classification performance a lot. However, the case of high the sigma and low the Danger represents the data space is little deformed by noise. The proposed model would not improve the data quality a lot.

b) *Binary ordinal classification with different rank r* : In this experiment, we set σ and m to be 1.5 and 10 respectively, and $X^{(k)}$ to have different ranks such as 1, 3, and 5. The purpose of this experiment is to show the robustness of SVDMC regardless of whether the data is a low-rank matrix or not.

According to Table 2, Table 5, and Table 6, SVDMC outperforms the other approaches for various values of r . Note that estimated values of the parameter β are almost constant for different ranks as shown in Table 7. This is because that β increases with the amount of noise in the data and hardly depends on the rank of the data. In the case of the same amount of noise, the rank of the data increased, and the amount of noise accumulated in the overlap section decreased (lower Danger values). That is, in the case of the same amount of noise, the proposed model performance may be better when the rank is small, but looking at Table 5, and Table 6 the rank itself is not significantly affected.

Table 5: Comparison of empirical classification results of synthetic binary data with $r = 3$

<i>wk</i> NN	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.53 ± 0.02	0.71 ± 0.04	0.56 ± 0.04	0.61 ± 0.05	0.56 ± 0.05	0.64 ± 0.05	0.51 ± 0.02
G-mean	0.25 ± 0.12	0.70 ± 0.05	0.42 ± 0.11	0.57 ± 0.08	0.37 ± 0.13	0.63 ± 0.07	0.12 ± 0.13

Table 6: Comparison of empirical classification results of synthetic binary data with $r = 5$

<i>wk</i> NN	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.51 ± 0.02	0.61 ± 0.05	0.52 ± 0.04	0.55 ± 0.04	0.53 ± 0.04	0.58 ± 0.05	0.50 ± 0.00
G-mean	0.14 ± 0.12	0.59 ± 0.06	0.24 ± 0.18	0.50 ± 0.06	0.29 ± 0.13	0.57 ± 0.06	0.04 ± 0.00

Table 7: Danger values and information of β_1 and β_2 of the second simulation.

	$r = 1$	$r = 3$	$r = 5$
Danger(original)	0.14	0.13	0.12
Danger(SVDMC)	0.07	0.06	0.05
β_1	30.94±1.74	29.90±1.84	29.86±1.80
β_2	7.04±2.27	6.79±2.69	8.89±2.43

Table 8: Comparison of empirical classification results of synthetic multiclass data

<i>wk</i> NN	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.51 ± 0.01	0.52 ± 0.03	0.51 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.50 ± 0.03	0.50 ± 0.00
G-mean	0.04 ± 0.06	0.33 ± 0.06	0.02 ± 0.06	0.30 ± 0.05	0.07 ± 0.09	0.35 ± 0.06	0.00 ± 0.00

c) Multiclass ordinal classification: We now compare the performance of the seven preprocessing methods in the ordinal classification problem with three classes. Let us denote the classes as Class 1, Class 2, and Class 3. Since most multiclass ordinal data do not have sufficient observations in classes with extreme levels, such as the lowest level or the highest level, we set the number of observations for Class 1 and Class 3 to 100, and for Class 2 to 1000. We also set the values of σ, r, m to 2, 1, 10 respectively. The estimated β values are (14.89, 48.73, 15.91) respectively. Under this setting, the proposed method still leads to promising results as shown in Table 8. Danger values decreased from 0.27 to 0.07 by the proposed model.

3.6 Real Data Example

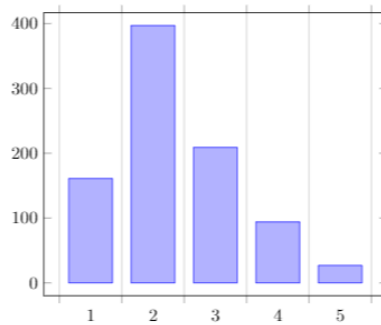


Figure 5: Distributions of the disaster data

The proposed method is applied to the real ordinal dataset collected from the Hungarian National Association of Radio Distress-Signaling and Infocommunications (RSOE) from May 2013 to December 2014. RSOE monitors emergency and disaster events happening all over the world in real time, displaying the information related to the events on its website and making it available to the public. All events are categorized separately into several types of disasters such as earthquakes, fire, floods, landslides, nuclear events, tornados, volcanic events, etc, with the information described in

Table 9: Experimental classification results of disaster data on $wkNN$ ($k=7$) algorithm with 7 different types of training sets

$wkNN$	Original	SVDMC	GAPCs	SMOTE	InCuBAte	SCUT	ENN
MAUC	0.87 ± 0.02	0.92 ± 0.01	0.87 ± 0.08	0.84 ± 0.02	0.82 ± 0.03	0.86 ± 0.03	0.79 ± 0.03
G-mean	0.80 ± 0.05	0.87 ± 0.02	0.81 ± 0.11	0.75 ± 0.04	0.65 ± 0.03	0.78 ± 0.04	0.63 ± 0.17

natural language. In addition, each event is categorized according to its damage level: nominal (level1), minor (level2), moderate (level3), severe (level4), or extreme (level5). The level of damage is a comprehensive risk index for the damage caused by a disaster and has been assessed and categorized by some disaster experts. Figure 5 depicts the distribution of the number of events belonging to each level. The data was originally introduced by [15] and we followed their procedure to build a term-document matrix from event descriptions. Note that the disaster data contain a lot data noise since their attributes (32 words in the term-document matrix) were built from a somehow subjective material. Nevertheless, Table 9 shows that our method outperforms other approaches and confirms the validity of using the SVDMC for imbalanced ordinal data. The mean of estimated parameters $\hat{\beta}$ as 26.69,32.58,35.56,34.07 and 17.49 for levels 1 to 5 respectively.

The calculation time per iteration was 0.16 min and there was an average of 46 iterations. The total calculation time was 7.36 in average. For comparison, a representative wrapper model for imbalanced data, the cost-sensitive algorithm based on C-Support Vector Classification was implemented [56]. The class weight is optimized by GAs in the same setup. The model had the calculation time 0.18 min per iteration, and it cannot converge to an optimal solution. Furthermore, the time complexity of the proposed model with Big O notation is $O(P \times I \times n \times (m+k))$ where P and I are the population size and number of iterations of GAs respectively, and n and m are the number of observations in a training set and the dimension of the data respectively, and k is the hyperparameter of the wkNN algorithm. The time complexity of GAs is proportional to the time complexity of a fitness function and number of fitness function evaluations, which is $O(P \times I)$. The time complexity of fitness function is then the sum of them of SVD, $O(n \times m)$, and evaluation of the classifier of the training data matrix, $O(n \times (m+k))$, so that the time complexity of the fitness function is $O(n \times (m+k))$. Therefore, the time complexity of SVDMC is $O(P \times I \times n \times (m+k))$ based on the classifier, wkNN algorithm for ordinal data.

4. Sensor Drift Compensation for Mixed Gas Classification under Batch Experiments

4.1 Introduction

An electronic nose (E-nose) is a semiconductor composed of an electronic chemical sensor array. It has been applied to the identification of individual components or gas mixtures [57, 58, 33], food industry quality control [59, 60, 61], public health [62, 63, 64], explosive detection [65] and a space program [66]. Especially it is very important to accurately classify mixed gases to reduce the risks of hazardous gases since most of the dangerous gases are mixed gases as referred in [67, 68, 69, 70]. Unfortunately, mixed gas classification is complicated because one type can contain different substances and concentrations.

Sensor drift under batch experiments is one of the well-known issues in mixed gas classification. Sensor drift indicates that a sensor has unexpected sensor values when exposed to complex environmental factors such as dynamic temperature, humidity, pressure, and poisoning. It is a major factor that makes it difficult to analyze the response of a sensor array and to perform gas recognition [71]. Drift can distort a unique pattern for the target odor resulting in poor representation of the target odors data. Furthermore, in the mixed gas, other chemical gases can generate an unexpected chemical reaction when exposed to the environmental factors. Therefore, it is essential to compensate drift for improving system reliability and increasing the accuracy of gas recognition through the sensor array system.

In this paper, we consider sensor array data detected over a short time, so we focus on drift due to the environmental covariates, such as temperature and humidity. This paper has the purpose of reconstructing drifted value based on a nonlinear parametric function of experimental covariates. The nonlinear model was constructed based on the correction model [72] that successfully compensated for the effects of ambient conditions. Also, we handle more demanding situations that are expensive to perform mixed gas experiments. Many studies for gas sensor analysis proceed with feature extraction such as steady state value extraction [73]. But in our difficult situation, machine learning faces overfitting because there is little data on gas analysis. To overcome this, we extract additional stable state response values are sampled from fixed environmental covariates so that machine learning technology can learn various response values under one environmental condition.

4.2 Literature Reviews

Much attention has been paid in recent years to address this issue, preprocessing techniques to deal

with sensor drift. These methods can be categorized into two categories: adaptive methods and feature transformation methods. There are many adaptive methods, for example, models based on Self-Organizing Maps (SOM) [74], Active Learning (AL) [75], and Deep Belief Network (DBN) [76], and the model applying classifier ensembles based on SVM [77], and an extreme learning machine for drift compensation [78] and more, which are passive models to handle sensor drift. They do not actively search for a basis occurring drift and only identify the slow changes of sensor responses caused by long-term drift. The performance of these models looks good, but they do not solve the drift issue using environmental covariates.

Most feature transformation techniques for drift are component calibration methods that regard the drift signal as divisible, so they aim to separate the drift signal from sensors signal. Feature transformation techniques for sensor drift conclude Component calibration methods include PCA [79], PCA-based Component Correction (PCA-CC) [80], Independent Component Analysis (ICA) [81], Partial Least Squares (PLS) [82], Orthogonal Signal Correction (OSC) [83], Discriminant Factorial Analysis (DFA) [84], Linear Discriminate Analysis (LDA) [78], Discriminative Domain Regularized Component Analysis (D-DRCA) [85], wavelet [86] etc. However, they only identify and analyze the effect of temperature and humidity on a specific sensor, not try to restore to the value before the drift occurred.

The model proposed in [87] tried to reconstruct drifted sensor responses based on the Papoulis-Gerchberg method. They suggest that the relationship between the measurements and the drifted signals is linear. They estimated the drifted signal that minimizes the errors in the reconstruction of the real sensor response based on the Lorentzian model. However, the paper focused on the drift effect as time goes, not fluctuating temperature and humidity values. The model proposed in [88] was used to eliminate drift effects due to ambient temperature and humidity fluctuations to obtain adequate precision in pollution level measurements. It adopts a nonlinear Multi-Input-Single-Output (MISO) system to construct an appropriate parametric structure. It was implemented by an Artificial Neural Network (ANN) to perform the transformation and describe the experimental results. The simple structure of ANN with one hidden layer was successful in the identification of different concentrations of Methanol. However, it is not enough to classify mixture gases under the existence of a few observations.

4.3 Methodology

We propose a preprocessing method for the drift compensation for accurate electronic nose data classification. In Sections 4.3.1, we introduce the correction model and our parametric model to handle the sensor drift. For the parameter tuning, in Section 4.3.2, we discuss a statistical function

and a heuristic algorithm for determining proper parameters for successful gas identification.

4.3.1 Sensor Drift Compensation Model

We assume that temperature and humidity are the main factors that greatly affect the function of sensor data. To compensate for the impact of the environmental covariates, we adopt the correction model to lead a drift compensation [72]. Let z_i be a vector of length n with the correction values with no drift for the sensor i , and it can be formulated as follows:

$$z_i = (\alpha_{i1}T_i + \alpha_{i2}RH_i + \alpha_{i3}\mathbf{1}_n) \circ V_{\text{Diff}_i} + \alpha_{i4}T_i + \alpha_{i5}RH_i \quad (4)$$

where T_i and RH_i are temperature and relative humidity vectors of length n , and the Hadamard product is described in \circ notation. V_{Diff_i} is a vector of length n with differences between active voltages and reference voltages for the sensor i . The active voltage responds to both target gas concentration and environmental covariates, while the reference voltage is affected only by the change of environmental covariates. And, α_{ij} are unknown parameters for $j = 1, \dots, 5$.

Since the sensors are independent, we apply our model to a sensor multiple times for multivariate analysis. To be specific, GAs learn temperature, relative humidity, and measurements V_{Diff_1} of sensor 1 to predict the parameters α_{1j} for $j = 1, \dots, 5$, and then we get the transformed data via Eq.(4). Next, GAs learn the data V_{Diff_2} of sensor 2 with the environmental covariates to obtain the five parameters α_{2j} for $j = 1, \dots, 5$, and then we get the transformed data. This process is repeated to other sensors.

The equation Eq.(4) contains the minimum parameters and represents the nonlinear relationship between environmental factors and voltages, which is suitable in the case of few observations. For reference, our objective is not to accurately formulate the correction value z_i , but to transform the observed data into a new data space with improved gas identification.

4.3.2 Class Separability Criterion of GAs

The goal is to estimate parameters that achieve sensor drift correction and improved classification performance both. We assume that different gases without drift should have different distributions and different data space areas. Thus, we set the fitness function for a genetic algorithm as a function evaluating how different classes differ in the data space. Note that improper fitness function would generate transformed data that lost its originality, so it is essential to determine the appropriate fitness features. Our fitness function was originated from the criteria of multiclass linear discriminant

analysis [89], which has similar objective as ours, finding orthogonal projections with the centers (averages) of different categories far from each other and with little variance.

In [89], the separability measure is defined as the ratio for the trace of the between class scatter matrix (S_B) and that of the within class scatter matrix (S_W). The expressions are as follow:

$$S_B = \sum_{c=1}^C n_c (m_c - m)(m_c - m)^T \quad (5)$$

$$S_W = \sum_{c=1}^C \sum_{j=1}^{n_c} (z_{c,j} - m_c)(z_{c,j} - m_c)^T \quad (6)$$

where $C \geq 2$ is the number of classes, and n_c and m_c are the number of individuals $z_{c,j}$ and the mean of the class c for $c = 1, \dots, C$. The variable m is the mean of the total observations. To prevent the impact of the imbalance issue, we normalized the equations Eq.(5) and Eq.(6) by the number of observations. Then, our criterion measure is by the ratio of the variance of classes' means S'_B to the sum of variances within classes S'_W where,

$$S'_B = \sum_{c=1}^C (m_c - m)^2$$

$$S'_W = \sum_{c=1}^C \frac{1}{n_c} \sum_{j=1}^{n_c} (z_{c,j} - m_c)^2$$

For scaling, we multiply the number of classes C to the ratio. Therefore, our objective function J is,

$$J = \left(\frac{S'_B}{S'_W} \right) \quad (7)$$

As we discussed earlier, GAs seek the optimized parameters in Eq.(7) maximizing the fitness function J . Figure 6 illustrates the pseudocode of the proposed drift compensation algorithm optimized by GAs.

Algorithm 2: The Algorithm of Drift Compensation

```

procedure DRIFT_COMPENSATION(Train=D,
Train.temp=T,Train.humi=RH,Test=DD,Test.temp=TT,Test.humi=HH)
2:    $s$  is the feature numbers (number of sensors) of input data  $D, DD$ 
    $n$  is the number of observations in  $D$ 
4:    $n_c$  is the number of experiments for odor  $c$ 
    $D_i, T_i, RH_i$  are vectors of length  $n$  for  $i = 1, \dots, s$ 
6:   procedure ALPHA( $\alpha_i = (\alpha_{i1}, \dots, \alpha_{i5})$ ) for  $i = 1, \dots, s$ 
    $z_i \leftarrow (\alpha_{i1}T_i + \alpha_{i2}RH_i + \alpha_{i3}\mathbf{1}_n) \circ D_i + \alpha_{i4}T_i + \alpha_{i5}RH_i$ 
8:    $m_c \leftarrow \frac{1}{n_c} \sum_{j=1}^{n_c} z_{c,j}$  for  $c = 1, \dots, C$ 
    $m \leftarrow \sum_{c=1}^C m_c$ 
10:   $S'_B \leftarrow \sum_{c=1}^C (m_c - m)^2$ 
    $S'_W \leftarrow \sum_{c=1}^C \frac{1}{n_c} \sum_{j=1}^{n_c} (z_{c,j} - m_c)^2$ 
12:  return  $J = \frac{S'_B}{S'_W}$ 
end procedure
14:   $\hat{\alpha} = \text{GA}(\text{ALPHA})$  ▷ Find  $\hat{\alpha}$  maximizing the value of the function ALPHA
    $\hat{z}_i \leftarrow (\hat{\alpha}_{i1}T_i + \hat{\alpha}_{i2}RH_i + \hat{\alpha}_{i3}\mathbf{1}_n)D_i + \hat{\alpha}_{i4}T_i + \hat{\alpha}_{i5}RH_i$  for  $i = 1, \dots, s$  ▷ Transformed training set
16:   $nn$  is the number of observations in  $DD$ 
    $\hat{z}_i \leftarrow (\hat{\alpha}_{i1}TT_i + \hat{\alpha}_{i2}HH_i + \hat{\alpha}_{i3}\mathbf{1}_{nn})DD_i + \hat{\alpha}_{i4}TT_i + \hat{\alpha}_{i5}HH_i$  for  $i = 1, \dots, s$  ▷ Transformed test set
18:  return  $\{\hat{z}, \hat{z}\}$ 
end procedure

```

Figure 6: Pseudocode of the proposed drift compensation algorithm

The parameter α can be solved by a spectrum decomposition of $S'_B^{-1}S'_W$. In the limited number of experiments, there often exists singularity to solve S'_B^{-1} , so the spectrum decomposition method cannot achieve an optimal solution. Therefore, a heuristic approach such as GAs is needed to find the optimal parameter α .

4.4 Description of Experimental Setup

The performance of our sensor drift compensation model is evaluated by comparing it with the existing correction model with ANN [88]. It constructed a nonlinear system to formulate the effect of sensor drift based on ANN. We call transformed data by our model and the ANN model by Correction set and ANN set respectively. However, in our knowledge, there is no filter model with a nonlinear structure to compare with the suggested model. As a classifier, we applied Random Forests (RF) [90], a Support Vector Machine (SVM) [91], K-nearest-neighbor (KNN) [92] and Multinomial Logistic Regression (MLR) [93].

For reference, we applied a genetic algorithm as follows. The crossover rate is 80%, and the mutation rate is 0.1. GAs run 100 generations with a population of 100. Evolution stops when there

has been no improvement over the past 30 generations.

In gas classification, the Misclassification Error Rate (MER) has been used, which is the basic concept in measuring and MER is the value calculated by the number of observations in a negative class divided by the total number of data. The range of MER is from 0 to 1, and closer to 1 indicated the better classification performance.

4.5 Simulation Study

The goal of this section is to confirm that the proposed model has the role of sensor drift compensation through simulation. In this simulation, we assume that different gases have different distributions of sensor values if there is no sensor drift. For simplicity, there are three types of odors, odor A, odor B, and odor C and five sensors; three of them detect the released gas, and the others measure temperature and relative humidity. The three gases, odor A, odor B, and odor C have different $\mathbf{z} = (z_1, z_2, z_3)^T$ values for sensor 1, sensor 2, and sensor 3. We generate \mathbf{z} values of sensor 1, sensor 2, and sensor 3 from uniform distributions of the intervals [220,250], [200,230], and [180,210] respectively. For example, odor A has $\mathbf{z} = (246, 208, 186)^T$, odor B has $(249, 211, 200)^T$ and odor C has $(250, 218, 207)^T$. Since there are five parameters per sensor, we denote the j th parameter of sensor i as α_{ij} for $i = 1, \dots, 3$ and $j = 1, \dots, 5$. The parameters $\alpha_{i1}, \alpha_{i2}, \alpha_{i4}$ and α_{i5} follow a uniform distribution of $[-1, 1]$ independently, and α_{i3} follows the same distribution from 0.1 to 1 for $i = 1, \dots, 3$. The range of the third ones, different from others, is to avoid singularities and to make that $V_{\text{Diff}_i} (= y_i)$ have a positive relation with z_i . To distinguish an estimated parameter through our method, these are called true parameters. We determine temperature T and relative humidity RH values from uniform distributions with the intervals [22°C, 28°C], [40%, 60%]. Then, z_i and control variables T_i and RH_i are taken into the inverse of

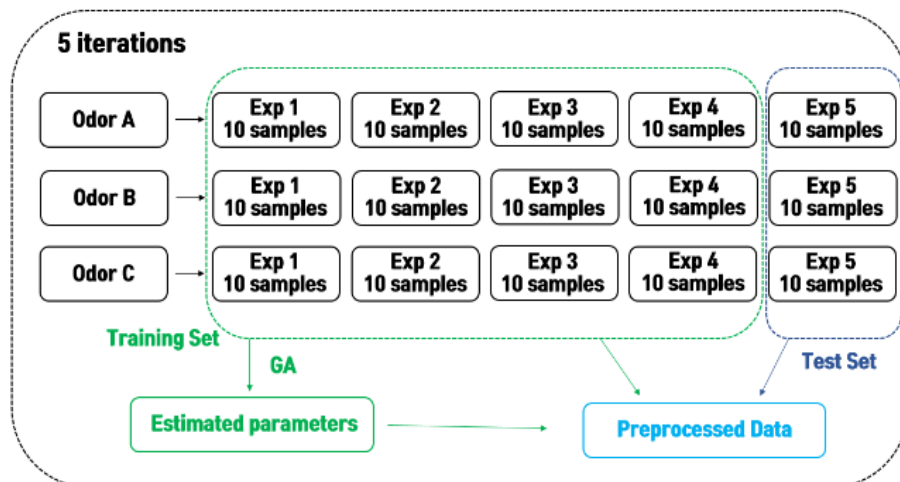


Figure 7: Block diagram of the simulation data preprocessing

Eq.(4) and add Gaussian noise of variance σ^2 ($\sigma = 0.1$ in our case) to get to the result the measurement y_i for $i = 1, \dots, 3$. We add noise to contain information about different drift effects and noise.

Since performing E-nose experiments are very costly, we consider the number of trials is limited. Therefore, we assume that there are only five experiments for each type of odor, which is too small to estimate the optimal parameters. To overcome this limitation, we apply a simple sampling technique to extract ten different y_i values under batch experiment. In this simulation, we generate ten different y_i values with different the amount of Gaussian noise. For real gas data, fluctuations in the measurements over time are used to complete sampling.

To train a classifier, four out of five experiments for each type of odor were set up as a training set (See Figure 7). We adjust the parameters with GAs using the training set and obtain the modified training set test set using the estimated parameters. By the sampling, a model can classify a test set into several types of odor. So, we determine the type of gas in the experiment as the type that was most frequently identified. This procedure is run a total of five times with different test sets to achieve unbiased classification performance.

Figure 8 illustrates the distributions of the generated data, \mathbf{z} and \mathbf{y} for sensor 1, sensor 2, and sensor 3. The values y_i represents sensor values affected by drift in five different environmental

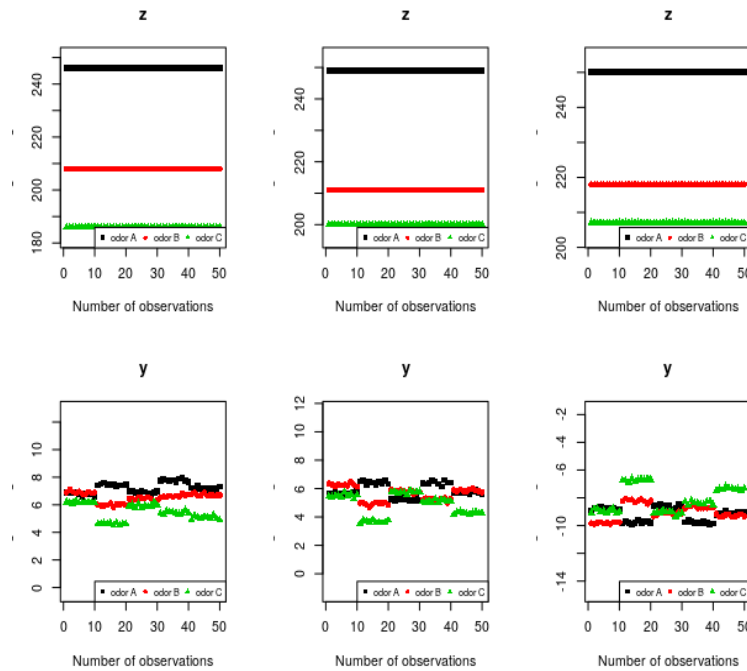


Figure 8: The distributions of simulation data; From left to right, figures represent the sensor values of sensor 1, sensor 2 and sensor 3.

Parameters	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}
True	0.46	0.39	0.16	0.04	0.32
Correction	0.64	0.53	0.50	-0.30	0.48

Table 10: Comparison among true parameters α_{1j} and the estimated ones $\hat{\alpha}_{1j}$ for $j = 1, \dots, 5$

conditions for $i = 1, \dots, 3$. We observe a large difference between the measured values for five experiments and some fluctuations caused by noise within one experiment. As expected, there is no variability within classes and no overlaps among classes in values \mathbf{z} . Conversely, it is hard to identify the distinct characteristics of odor A, odor B, and odor C from the values \mathbf{y} . They have a high variability in one class, and there is a large area of overlap among classes. We reconstruct the sensor response $\hat{\mathbf{z}}$ using the estimated parameters by Eq.(4).

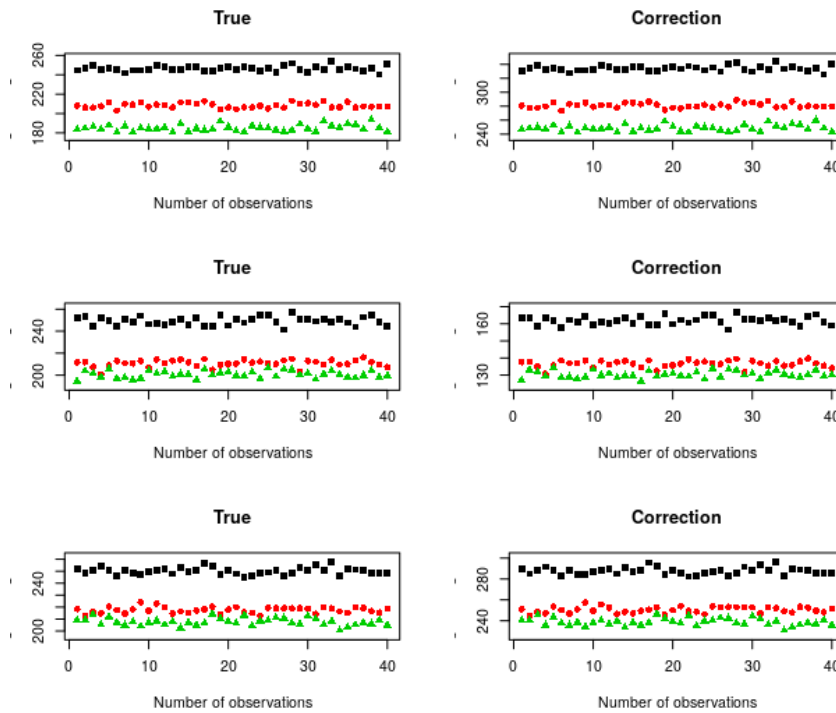


Figure 9: The distributions of simulation data; From left to right, figures represent transformed training data by true parameters and by the suggested correction model.

	RF	SVM	KNN	MLR
Original	0.42 ± 0.25	0.47 ± 0.31	0.46 ± 0.19	0.80 ± 0.11
True	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.03
Correction	0.99 ± 0.02	0.96 ± 0.07	0.95 ± 0.09	0.99 ± 0.01

Table 11: Result of classifications. The average of classification performances on ANN set was 0.41.

Our goal is to find patterns of sensor values that are not affected by drift. Table 10 represents true parameters (True) and estimated parameters (Correction) by GAs. Note that true parameters are not the optimal parameters for gas classification due to the Gaussian noise. Therefore, it is much valuable to check the similarity, not accuracy. Especially, the fourth and fifth parameters are dependent on temperature and relative humidity respectively and are independent of measurement \mathbf{y} . It is only necessary to confirm that the linear combination with temperature and relative humidity by the corresponding coefficients is similar with the combination by true parameters. In the table, it has been confirmed that the estimated parameters have a similar pattern to the true parameters so that the drift effect can be compensated. This can also be seen in the plots of the distributions of $\hat{\mathbf{z}}$ in Figure 9. The similarity between a reconstructed distribution with true parameters and a transformed distribution with estimated ones shows the model's ability to compensate for drift. The transformed data confirms that the suggested model can compensate simulated drift successfully.

Table 11 lists the contingency table of classification performances: Original is classification results of original data and the other performances on transformed data are named as the type of parameters. According to the table, the classifiers cannot accurately identify even half of the odors when data is affected by drift. On the other hand, all preprocessed data sets improved performance. As we expected, the reconstructed data by true parameters had suitable distributions to classify. Correction data sets also achieved performance as good as True. The average of classification performances on ANN set was 0.41, which there was no improvement on mixed gas identification. Therefore, the simulation data confirmed the superior performance of our model in handling drift and gas discrimination.

4.6 Real Data Example

The proposed method is applied to a real electronic nose data collected from the UCI machine learning repository [94]. The set-up system used in the experiment detected the generated the desired concentrations of C_2H_4 , CO , CH_4 , $\text{C}_2\text{H}_4 + \text{CO}$ mixtures and $\text{C}_2\text{H}_4 + \text{CH}_4$ mixtures. Each

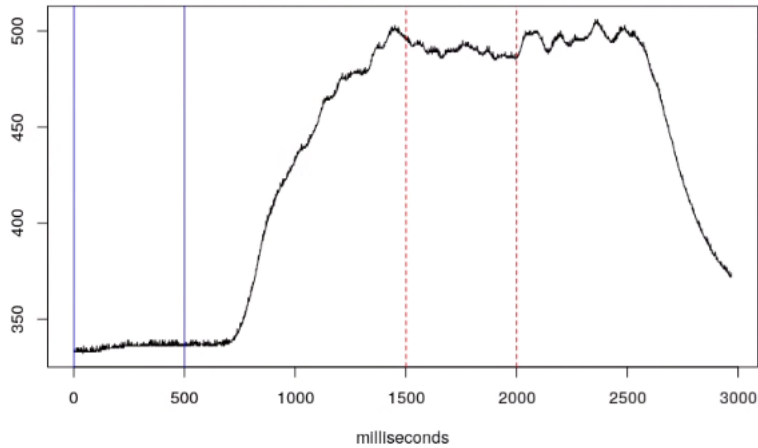


Figure 10: A gas is measured during 3000ms. Reference voltages and effective voltages along the areas marked with blue solid and red dotted lines respectively

volatile was released in four different flows providing 30 different mixture configurations: C_2H_4 (0ppm, 31ppm, 46ppm, 96ppm), CO (0ppm, 270ppm, 397ppm, 460ppm), CH_4 (0ppm, 51ppm, 115ppm, 131ppm). The E-nose contains eight MOX gas sensors that produce time-dependent responses to the different gas stimuli. Temperatures and relative humidity are measured either. Each measurement had a total duration of 300 seconds, measured every 100ms. During the first 60 seconds, no gas was released from the gas sources, which corresponds to the reference voltage. In 60 seconds, the sources begin to release the gas mixtures at the specified flow rate. The detected sensor values quickly change to a steady state that corresponds to the active voltage as shown in Figure 10. On the other hand, we observed some abnormal sensor values. In the case of a small number of observations, ML does not properly operate if outliers exist. For better gas classification, we removed outliers and applied our methodology and classifiers. To identify outliers, we used a Local Outlier Factor (LOF) that gives the degree of being an outlier of observation [95]. The degree depends on how isolated the observation is from the surroundings. We select the value of MinPts, the nearest neighbors, as three. We removed nine experiments with values greater than twice the standard deviation from the mean. We analyze the discriminative properties of the different concentration levels of C_2H_4 . It was not possible to discriminate the concentration levels of C_2H_4 using only the steady-state response. Figure 11 shows the steady state values of sensor 1 by the group. Et_H represents a group of mixed gas observations with high ethanol concentrations. Et_M and Et_L are with medium and low ethanol concentrations respectively. Observations of mixed gas without ethanol are Et_n . As the ethanol concentration of the mixed gas decreases, the sensor value is generally reduced. However, the overlap regions between classes are very large so it is hard to identify distinct feature of groups.

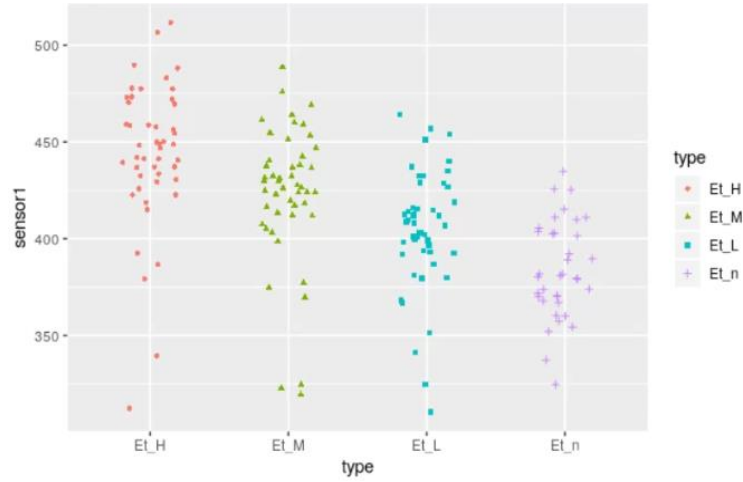


Figure 11: A scatter plot showing the steady state values of sensor 1 by the group.

	RF	SVM	KNN	MLR
Original	0.66 ± 0.10	0.74 ± 0.10	0.64 ± 0.10	0.73 ± 0.11
Correction	0.78 ± 0.08	0.79 ± 0.08	0.69 ± 0.09	0.77 ± 0.09

Table 12: Summary of classification performance

nearest neighbors, as three. We removed nine experiments with values greater than twice the standard deviation from the mean. We analyze the discriminative properties of the different concentration levels of C_2H_4 . It was not possible to discriminate the concentration levels of C_2H_4 using only the steady-state response. Figure 11 shows the steady state values of sensor 1 by the group. Et_H represents a group of mixed gas observations with high ethanol concentrations. Et_M and Et_L are with medium and low ethanol concentrations respectively. Observations of mixed gas without ethanol are Et_n . As the ethanol concentration of the mixed gas decreases, the sensor value is generally reduced. However, the overlap regions between classes are very large so it is hard to identify distinct feature of groups.

As we discussed earlier, we get V_{Diff_i} by a difference between of active voltage and reference voltage. Since there are some fluctuations in measurements, we extracted 100 measurements from ranges corresponding to the active voltage and reference voltage respectively and averaged those samples. For a reference voltage, we took randomly 100 measurements from 1 to 500ms as shown by the blue solid lines in Figure 10. To get an active voltage, we took randomly 100 measurements from 1501 to

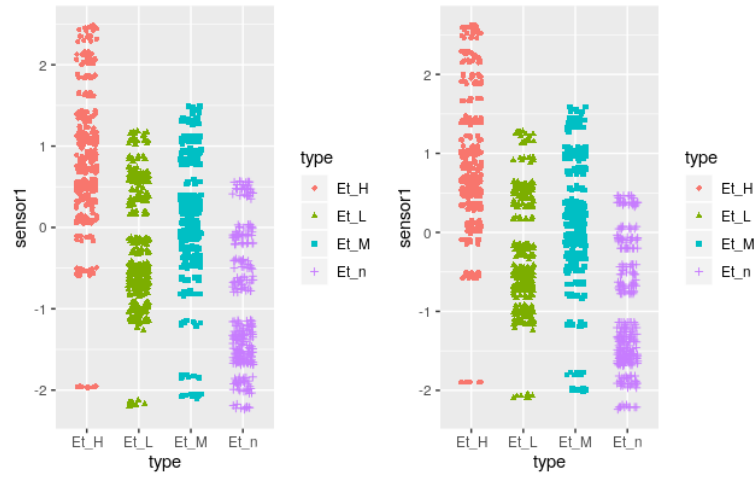


Figure 12: Scatter plots of original sensor 1 values and corrected sensor 1 values from left to right

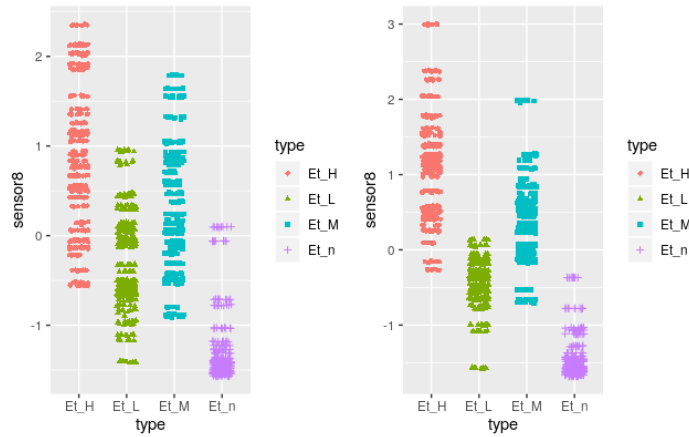


Figure 13: Scatter plots of original sensor 8 values and corrected sensor 8 values from left to right

2000ms as shown by the red dotted lines in the figure. In addition, we used one temperature value and one humidity value that are the average values of temperature and humidity measured for 300 seconds. Like the simulation, the sampling was introduced so that the data consisted of ten different V_{Diff_i} under batch experiments for $i = 1, \dots, 8$. This process allows us to have various sensor values without losing the properties of the gas sensor values of the experiment.

The proposed method confirms that the modified dataset for drift improved the classification performance of all classifiers (See Table 12). The proposed method enhanced the performance of classification up to 18.18%. The classification accuracy on ANN set was about 0.44. It is obvious because it requires many observations to predict parameters.

In the UCI data, eight sensors independently detect gas signals so that different sensors have different

effects on temperature and relative humidity. Not all of them are affected by covariates, so some of gas arrays were hardly transformed and some were transformed properly. Specifically, the values of sensor 1, 2, 3, and 4 did not show significant improvement in reducing within-variance and overlap regions among classes. Figure 12 represents two scatter plots of sensor 1 from original data set and correction data set from left to right. For comparison, we proceed a scale transformation on the values. We confirmed that the parameters $\hat{\alpha}_{11}$ and $\hat{\alpha}_{12}$ were close to zero, which means that the corrected value z_1 has meaningful relationship with only $V_{\text{Diff}1}$. On the other hand, corrected sensor values from sensor 5, 6, 7 and 8 showed significant reduction in within-variance and overlap regions. Figure 13 describes the results of our model regrading sensor 8. Thus, the UCI data seems to have only half of sensors that are affected by environmental covariates. When more sensors affected by temperature and humidity; the degree of improvement will increase.

5. Conclusion

5.1 Summary and Contributions

In supervised learning, the impact of feature transformation is very significant. Transformed features can produce more concise and accurate information. The transformed features exist in a new space with higher data quality and better separability. This paper proposed two feature transformation techniques for ordinal-imbalanced data and sensor array data with drift. SVDMC reduces the additive noise in ordinal data resulting in making a balance in information among different classes. The sensor drift compensation for mixed gas classification under batch experiments successfully corrects for the effect of environmental covariates on sensor arrays. They improve classification performance by solving imbalance and sensor drift problems, respectively. The empirical results on simulation datasets and real datasets showed consistent improvement in learning performance.

We proposed a novel preprocessing method called SVDMC to diminish the effect of class imbalance problems in ordinal data multiclass classification. The proposed method solves the imbalance problem by reducing data noise and the class overlap area, without modifying the distribution of observations over multiple classes. In this paper, we argue that the key to solving the imbalance issue in ordinal data is not to adjust the number of data points over classes but to improve the data quality by reducing the noise of the data. From the results of numerical experiments, the proposed method generally outperformed other existing approaches based on two different accuracy measures: MAUC and G-mean.

Gas sensors drift generates serious limitations in the appliance and development of electronic noses. The ideal operating state of the electronic nose is to maintain the same response whenever the same volatile compound comes on the gas sensor surface. But in the real appliance, sensors are aging and being affected by various environmental factors. This results in sensor drift that significantly reduces the ability of the electronic nose. In this paper, in order to compensate for the drift, we reconstruct sensor values based on the nonlinear parametric compensation model. The parameters are determined by values that optimize a well-defined objective function. We propose an objective function with greater value as the variance of individual groups becomes smaller and the difference between different classes becomes larger. It helps to correct the sensor drift caused by the environmental covariates, but also substantially improves classification performance.

5.2 Limitations and future research

In our future research, we will analyze the behavior of SVDMC when combined with various ordinal

data classifiers. Moreover, based on the literature on matrix denoising [96], solving the optimal shrinkers for ordinal classification of imbalanced data would be considered for future research. Developing the proposed model with a filter method reduces calculation time and makes it easier to interpret the results.

In addition, for the sensor drift compensation model, we desire to create a mixture model based on our method. It would be more effective because the drift effect varies depending on the type of gas. In this study, we attempted to infer a pattern of corrected sensor values, but estimating specific values using chemical information is considered for future research.

References

- [1] S. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111-117, 2006.
- [2] M. Prakash and M. Narasimha Murty, "A genetic approach for selection of (near-) optimal subsets of principal components for discrimination," *Pattern Recognition Letters*, vol. 16, no. 8, pp. 781-787, 1995.
- [3] P. Sánchez, F. G. Montoya, F. Manzano-Agugliaro and C. Gil, "Genetic algorithm for S-transform optimisation in the analysis and classification of electrical signal perturbations," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6766-6777, 2013.
- [4] Z. Huang, M. Pei, E. D. Goodman, Y. Huang and G. Li, "Genetic Algorithm Optimized Feature Transformation—A Comparison with Different Classifiers," in *Genetic and Evolutionary Computation Conference*, 2003.
- [5] G. Zames, N. Ajlouni, J. Holland, W. Hills and D. Goldberg, "Genetic algorithms in search, optimization and machine learning.," *Information Technology Journal*, vol. 3, no. 1, pp. 301-302, 1981.
- [6] N. A. Abolkarlou, A. A. Niknafs and M. K. Ebrahimpour, "Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm," in *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2014.
- [7] G. Stein, B. Chen, A. S. Wu and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," in *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, 2005.
- [8] T.-S. Li, "Feature selection for classification by using a GA-based neural network approach," *Journal of the Chinese Institute of Industrial Engineers*, vol. 23, no. 1, pp. 55-64, 2006.
- [9] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert Systems with applications*, vol. 3, no. 2, pp. 231-240, 2006.
- [10] C. D. Stefano, F. Fontanella, C. Marrocco and A. Scotto di Fr, "A GA-based feature selection approach with an application to handwritten character recognition," *Pattern Recognition Letters*, vol. 35, pp. 130-141, 2014.
- [11] A. Meyer-Baese and V. J. Schmid, "Genetic Algorithms," in *Pattern recognition and signal analysis in medical imaging*, Elsevier, 2014, pp. 135-149.
- [12] M. L. Raymer, W. F. Punch, E. Goodman, L. A. Kuhn and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE transactions on evolutionary computation*, vol. 4, no. 2, pp. 164-171, 2000.
- [13] D. E. Courte, M. M. Rizki, L. A. Tamburino and R. Gutierrez-Osuna, "Evolutionary optimization of Gaussian windowing functions for data preprocessing," *International Journal on Artificial*

Intelligence Tools, vol. 12, no. 1, pp. 17-35, 2003.

- [14] A. Agresti, *Analysis of ordinal categorical data*, New York : John Wiley & Sons, 2019.
- [15] S. Kim, H. Kim and Y. Namkoong, "Ordinal classification of imbalanced data with application in emergency and disaster information services," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 50-56, 2016.
- [16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering*, no. 9, pp. 1263-1284, 2008.
- [17] I. Domingues, J. P. Amorim, P. H. Abreu, H. Duarte and J. Santos, "Evaluation of oversampling data balancing techniques in the context of ordinal classification," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.
- [18] F. C. Marques, H. Duarte, J. A. M. Santos, I. Domingues, J. P. Amorim and P. H. Abreu, "An iterative oversampling approach for ordinal classification," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019.
- [19] M. Pérez-Ortiz, P. A. Gutiérrez, C. Hervás-Martínez and X. Yao, "Graph-based approaches for over-sampling in the context of ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1233-1245, 2015.
- [20] I. Nekooimehr and S. K. Lai-Yuen, "Cluster-based weighted oversampling for ordinal regression (CWOS-Ord)," *Neurocomputing*, vol. 218, pp. 51-60, 2016.
- [21] I. Nekooimehr and S. K. Lai-Yuen, "Adaptive semi-supervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Systems with Applications*, vol. 46, pp. 405-416, 2016.
- [22] M. Pérez-Ortiz, A. Sáez, J. Sánchez-Monedero, P. Gutiérrez and C. Hervás-Martínez, "Tackling the ordinal and imbalance nature of a melanoma image classification problem," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016.
- [23] R. Cruz, K. Fernandes, J. F. Pinto Costa, M. P. Ortiz and J. S. Cardoso, "Ordinal class imbalance with ranking," in *Iberian conference on pattern recognition and image analysis*, 2017.
- [24] M. Dorado-Moreno, M. Pérez-Ortiz, M. D. Ayllón-Terán, P. A. Gutiérrez and C. H. Martínez, "Ordinal evolutionary artificial neural networks for solving an imbalanced liver transplantation problem," in *International Conference on Hybrid Artificial Intelligence Systems*, 2016.
- [25] L. Jiang, C. Qiu and C. Li, "A novel minority cloning technique for cost-sensitive learning," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 4, p. 1551004, 2015.
- [26] Y. Sun, M. S. Kamel, A. K. Wong and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.
- [27] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119-1130, 2012.

- [28] G. E. Batista, R. C. Prati and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20-29, 2004.
- [29] T.-M. Chang and W.-F. Hsiao, "Model-based collaborative filtering to handle data reliability and ordinal data scale," in *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2011.
- [30] J. V. Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data & Knowledge Engineering*, vol. 68, no. 12, pp. 1513-1542, 2009.
- [31] D. Devi, S. K. Biswas and B. Purkayastha, "Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique," *Connection Science*, pp. 1-38, 2019.
- [32] S. Puthusserypady and D. Narayana Dutt, "SVD based technique for noise reduction in electroencephalographic signals," *Signal Processing*, vol. 55, no. 2, pp. 179-189, 1996.
- [33] S. K. Jha and R. D. S. Yadava, "Denoising by singular value decomposition and its application to electronic nose data processing," *IEEE Sensors Journal*, vol. 11, no. 1, pp. 35-44, 2010.
- [34] J. J. Jena, G. Girish and M. Patro, "Evaluating Effectiveness of Color Information for Face Image Retrieval and Classification Using SVD Feature," in *International Conference on Advances in Computing and Data Sciences*, 2018.
- [35] W. A. Sethares, A. Ingle, T. Krč and S. Wood, "Eigentextures: An SVD approach to automated paper classification," in *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, 2014.
- [36] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408-421, 1972.
- [37] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.
- [38] J. Stefanowski and S. Wilk, "Improving rule based classifiers induced by MODLEM by selective pre-processing of imbalanced data," in *Proc. of the RSKD Workshop at ECML/PKDD, Warsaw*, 1997.
- [39] K. Napierała, J. Stefanowski and S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples," in *International Conference on Rough Sets and Current Trends in Computing*, 2010.
- [40] H. Han, W.-Y. Wang and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*, 2005.
- [41] C. Bunkhumpornpat, K. Sinapiromsaran and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Pacific-Asia conference on knowledge discovery and data mining*, 2009.

- [42] J. A. Sáez, J. Luengo, J. Stefanowski and F. Herrera, "SMOTE--IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences*, vol. 291, pp. 184-203, 2015.
- [43] A. Agrawal, H. L. Viktor and E. Paquet, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling," in *KDIR 2015 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)*, 2015.
- [44] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *The Annals of Statistics*, vol. 46, no. 1, pp. 60-89, 2018.
- [45] K. Konstantinides and K. Yao, "Statistical analysis of effective singular values in matrix rank determination," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 5, pp. 757-763, 1988.
- [46] G. W. Stewart, "Perturbation theory for the singular value decomposition," in *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, 1991.
- [47] H. Golub and C. F. Van Loan, "Matrix computations," Press, London, 1996.
- [48] D. L. Donoho, "De-noising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613-627, 1995.
- [49] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the american statistical association*, vol. 90, no. 432, pp. 1200-1224, 1995.
- [50] D. L. Donoho, I. M. Johnstone, G. Kerkycharian and D. Picard, "Wavelet shrinkage: asymptopia?," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 2, pp. 301-337, 1995.
- [51] K. Hechenbichler and K. Schliep, "Weighted k-nearest-neighbor techniques and ordinal classification," *Discussion Paper 399, SFB*, 2004.
- [52] A. Sanz-Garcia, J. Fernandez-Ceniceros, F. Antonanzas-Torres, A. Pernia-Espinoza and F. Martinez-de-Pison, "GA-PARSIMONY," *Appl. Soft Comput.*, vol. 35, no. C, pp. 1568-4946, 2015.
- [53] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119-1130, 2012.
- [54] Y. Sun, M. S. Kamel and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006.
- [55] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171-186, 2001.
- [56] C.-C. Chang and C.-J. Lin, "{LIBSVM}: A library for support vector machines," *ACM*

Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 27:1-27:27, 2011.

- [57] J. W. Gardner and P. N. Bartlett, "A brief history of electronic noses," *Sensors and Actuators B: Chemical*, vol. 13, no. 1, pp. 210-211, 1994.
- [58] F. J. R.-L. I, T. M, V. A and H. R, "Chemical discrimination in turbulent gas mixtures with mox sensors validated by gas chromatography-mass spectrometry," *Sensors*, vol. 14, no. 10, pp. 19336-19353, 2014.
- [59] T. Konduru, G. C. Rains and C. Li, "Detecting sour skin infected onions using a customized gas sensor array," *Journal of Food Engineering*, vol. 160, pp. 19-27, 2015.
- [60] Y. Dai, R. Zhi, L. Zhao, H. Gao, B. Shi and H. Wang, "Longjing tea quality classification by fusion of features collected from E-nose," *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 63-70, 2015.
- [61] V. Y. Musatov, V. Sysoev, M. Sommer and I. Kiselev, "Assessment of meat freshness with metal oxide sensor microarray electronic nose: A practical approach," *Sensors and Actuators B: Chemical*, vol. 144, no. 1, pp. 99-103, 2010.
- [62] D. Li, T. Lei, S. Zhang, X. Shao and C. Xie, "A novel headspace integrated E-nose and its application in discrimination of Chinese medical herbs," *Sensors and Actuators B: Chemical*, vol. 221, pp. 556-563, 2015.
- [63] Y. Adiguzel and H. Kulah, "Breath sensors for lung cancer diagnosis," *Biosensors and Bioelectronics*, vol. 65, pp. 121-138, 2015.
- [64] O. Burfeind, M. Bruins, A. Bos, I. Sannmann, R. Voigtsberger and W. Heuwieser, "Diagnosis of acute puerperal metritis by electronic nose device analysis of vaginal discharge in dairy cows," *Theriogenology*, vol. 82, no. 1, pp. 64-70, 2014.
- [65] A. Norman, F. Stam, A. Morrissey, M. Hirschfelder and D. Enderlein, "Packaging effects of a novel explosion-proof gas sensor," *Sensors and Actuators B: Chemical*, vol. 95, no. 1, pp. 287-290, 2003.
- [66] R. C. Young, W. J. Buttner, B. R. Linnell and R. Ramesham, "Electronic nose for space program applications," *Sensors and Actuators B: Chemical*, vol. 93, no. 1, pp. 7-16, 2003.
- [67] A. Somov, A. Karelin, A. Baranov and S. Mironov, "Estimation of a gas mixture explosion risk by measuring the oxidation heat within a catalytic sensor," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 12, pp. 9691-9698, 2017.
- [68] S. Kiani, S. Minaei and M. Ghasemi-Varnamkhasti, "Application of electronic nose systems for assessing quality of medicinal and aromatic plant products: A review," *Journal of Applied Research on Medicinal and Aromatic Plants*, vol. 3, no. 1, p. 2016, 1-9.
- [69] A. Loutfi, S. Coradeschi, G. K. Mani, P. Shankar and J. B. B. Rayappan, "Electronic noses for food quality: A review," *Journal of Food Engineering*, vol. 144, pp. 103-111, 2015.
- [70] A.-C. Romain and J. Nicolas, "Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview," *Sensors and Actuators B: Chemical*, vol. 146, no. 2,

p. 2010, 502-506.

- [71] A. Hierlemann and R. Gutierrez-Osuna, "Higher-order chemical sensing," *Chemical reviews*, vol. 108, no. 2, pp. 563-613, 2008.
- [72] P. Wei, Z. Ning, S. Ye, L. Sun, F. Yang, K. C. Wong, D. Westerdahl and P. K. Louie, "Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring," *Sensors*, vol. 18, no. 2, p. 59, 2018.
- [73] B. Mumyalmaz, A. Özmen, M. Ali Ebeoğlu, C. Taşaltın and İ. Gürol, "A study on the development of a compensation method for humidity effect in QCM sensor responses," *Sensors and Actuators B: Chemical*, vol. 147, no. 1, pp. 277-282, 2010.
- [74] S. D. Carlo, M. Falasconi, E. Sánchez, A. Scionti, G. Squillero and A. Tonda, "Exploiting evolution for an adaptive drift-robust classifier in chemical sensing," in *European Conference on the Applications of Evolutionary Computation*, 2010.
- [75] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen and D. Tao, "Cost-sensitive feature selection by optimizing f-measures," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1323-1335, 2017.
- [76] T. Liu, D. Li, J. Chen, Y. Chen, T. Yang and J. Cao, "Active Learning on Dynamic Clustering for Drift Compensation in an Electronic Nose System," *Sensors*, vol. 19, no. 16, p. 3601, 2019.
- [77] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," *Sensors and Actuators B: Chemical*, vol. 166, pp. 320-329, 2012.
- [78] L. Zhang and D. Zhang, "Domain adaptation extreme learning machines for drift compensation in E-nose systems," *IEEE Transactions on instrumentation and measurement*, vol. 64, no. 7, pp. 1790-1801, 2014.
- [79] A. Ziyatdinov, S. Marco, A. Chaudry, K. Persaud, P. Caminal and A. Perera, "Drift compensation of gas sensor array data by common principal component analysis," *Sensors and Actuators B: Chemical*, vol. 146, no. 2, p. 2010, 460-465.
- [80] T. Artursson, T. Eklöv, I. Lundström, P. Mårtensson, M. Sjöström and M. Holmberg, "Drift correction for gas sensors using multivariate methods," *Journal of chemometrics*, vol. 14, no. 5-6, pp. 711-723, 2000.
- [81] M. Kermit and O. Tomic, "Independent component analysis applied on gas sensor array measurement data," *IEEE Sensors Journal*, vol. 3, no. 2, pp. 218-228, 2003.
- [82] R. Gutierrez-Osuna, "Drift reduction for metal-oxide sensor arrays using canonical correlation regression and partial least squares," *Electronic Noses and Olfaction*, pp. 147-152, 2000.
- [83] M. Padilla, A. Perera, I. Montoliu, A. Chaudry, K. Persaud and S. Marco, "Drift compensation of gas sensor array data by orthogonal signal correction," *Chemometrics and Intelligent Laboratory Systems*, vol. 100, no. 1, pp. 28-35, 2010.
- [84] C. Delpha, M. Lumbreras and M. Siadat, "Discrimination and identification of a refrigerant gas in a humidity controlled atmosphere containing or not carbon dioxide: application to the electronic

- nose," *Sensors and Actuators B: Chemical*, vol. 98, no. 1, pp. 46-53, 2004.
- [85] Z. Yi and C. Li, "Anti-drift in electronic nose via dimensionality reduction: a discriminative subspace projection approach," *IEEE Access*, vol. 7, pp. 170087-170095, 2019.
- [86] L. Jun-hua, S. Zhong-ru and D. hui, "Drift reduction of gas sensor by wavelet and principal component analysis," *Sensors and Actuators B: Chemical*, vol. 96, no. 1-2, pp. 354-363, 2003.
- [87] D. Huang and H. Leung, "Reconstruction of drifting sensor responses based on papoulis-gerchberg method," *IEEE Sensors Journal*, vol. 9, no. 5, pp. 595-604, 2009.
- [88] F. Hossein-Babaei and V. Ghafarinia, "Compensation for the drift-like terms caused by environmental fluctuations in the responses of chemoresistive gas sensors," *Sensors and Actuators B: Chemical*, vol. 143, no. 2, pp. 641-648, 2010.
- [89] K. T. Abou-Moustafa, F. De La Torre and F. P. Ferrie, "Pareto models for discriminative multiclass linear dimensionality reduction," *Pattern Recognition*, vol. 48, no. 5, pp. 1863-1877, 2015.
- [90] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [91] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992.
- [92] E. Fix, Discriminatory analysis: nonparametric discrimination, consistency properties, USAF school of Aviation Medicine, 1951.
- [93] D. Hosmer and S. Lemeshow, "The multinomial logistic regression model," *Hosmer D, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons*, pp. 260-287, 2000.
- [94] D. Dua and C. Graff, "UCI Machine Learning Repository," URL <http://archive.ics.uci.edu/ml>, 2017.
- [95] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000.
- [96] M. Gavish and D. L. Donoho, "Optimal shrinkage of singular values," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2137-2152, 2017.

Acknowledgements

Foremost, I would like to express my special thanks of gratitude to my academic advisor Professor, Sungil Kim for the continuous support of my master's degree and research, for constant encouragement and overall support for a meaningful academic life. He gave me a lot of opportunities to have many valuable expertise and study these interesting topics. I wouldn't reach this point without his presence.

Besides my advisor, I would like to thank to my committee for their continued support and encouragement: Professor, Chiehyeon Lim and Professor, Junghye Lee. Thanks to their many stimulating questions, the project was able to complete successfully. Also, meetings and conversations with them were very important in encouraging me to think of comprehensive and objective criticism from multiple perspectives.

I would also like to thank the following people, without whom I would not have been able to complete this research, and without whom I would not have made it through my master's degree: Thanks to every member of DA lab and the 209 classmates for their support and encouragement. They guided me so positively and always made me feel confident in my abilities. And for my love Yun and family, thanks for all their constant love and support. I simply couldn't have done this without them, special thanks.

