

Combined Features with Global and Local Features for Gas Classification

Sang-Il Choi *

Abstract

In this paper, we propose a gas classification method using combined features for an electronic nose system that performs well even when some loss occurs in measuring data samples. We first divide the entire measurement for a data sample into three local sections, which are the stabilization, exposure, and purge; local features are then extracted from each section. Based on the discrimination analysis, measurements of the discriminative information amounts are taken. Subsequently, the local features that have a large amount of discriminative information are chosen to compose the combined features together with the global features that extracted from the entire measurement section of the data sample. The experimental results show that the combined features by the proposed method gives better classification performance for a variety of volatile organic compound data than the other feature types, especially when there is data loss.

▶ Keyword : Gas classification, Electronic nose, Discriminant analysis, Global feature, Local feature, Feature selection, Combined feature

I. Introduction

전자코(electronic nose)는 센서를 이용하여 특정한 기체의 종류를 분류하는 전자 시스템이다[1,2,3]. 인간의 후각 기능은 쉽게 피로해져서 연속적으로 다른 냄새를 맡지 못하고 맡을 수 있는 냄새의 종류가 한정되어 있는 반면, 전자코는 지속적으로 냄새를 모니터링 할 수 있으며 인간에 유해한 기체에 대해서도 분간할 수 있는 장점이 있다[4,5]. 초창기의 전자코 시스템은 열량 측정(calorimetric) 센서를 이용하여 기체의 측정치를 색 배열(array of color)의 형태로 표현하였다[6]. 이러한 시스템은 기체를 분류하기 위해 가스 크로마토그래피(gas chromatography) 시스템이나 복잡한 MI(Machine Intelligence)가 결합된 질량 분석계(mass spectrometer)와 같은 정밀 장비를 포함하여 여러 복잡한 분석 과정이 요구되기 때문에 사용 환경에 제약이 있어 활용 분야에 있어서도 음식, 음료 분야의 품질 관리 등에 한정되었다[7,8]. 그러나 전자 화학 센서(electrochemical sensor)와 하드웨어 및 소프트웨어 IT 기술의 발달함에 따라 전자코 시스템이 보다 능동적이고 휴대가 가능해지면서 의료 검진용, 오염 물질 및 가스 누출 검색,

환경 보호, 화장품 산업에 이르기까지 전자코의 응용 분야는 점점 더 다양해지고 있다[9,10,11].

전자코 시스템은 여러 개의 채널로 구성된 센서 어레이(sensor array)와 컴퓨팅(computing) 시스템으로 구성되어 있다[10]. 센서의 각 채널은 각각 다른 특성을 가지고 있어 기체의 종류에 따라 각각 다르게 반응한다. 센서의 반응은 전자적인 인터페이스에 의해 벡터 형태의 수치 데이터로 변환된다. 컴퓨팅 시스템에서는 이렇게 변환된 데이터로부터 기체 분류에 유용한 특징을 추출하고, 분류기(classifier)를 통해 기체의 종류를 분별한다. 따라서 센서 어레이 감도(sensitivity)와 특징추출 및 분류기의 성능은 전자코 시스템의 성능을 결정하는 매우 중요한 요소이다.

전자코 시스템에 사용되는 다양한 유형의 센서 어레이들 가운데, 전도성 고분자 복합물질(conducting polymer composites), intrinsically 전도성 고분자(conducting polymer), 금속 산화물(metal oxides)은 전도율 센서(conductivity sensor)들에서 물질을 감지 할 때 많이 사용된다[12,13,14]. 고분자 복합물질은 센서 표류(sensor drift), 센

• First Author: Sang-Il Choi, Corresponding Author: Sang-Il Choi

*Sang-Il Choi(choisi@dankook.ac.kr), Dept. of Computer Science and Engineering, Dankook University

• Received: 2015. 11. 09, Revised: 2015. 11. 23, Accepted: 2015. 11. 30.

서 수명의 제한, 온도나 습도에 대한 민감성 등의 단점이 있음에도 불구하고, 광범위한 종류의 복합 물질들에 사용될 수 있고 비용이 많이 들지 않으며 실내 온도에서의 안정적인 동작, 적은 전력 소모 등, 다른 물질들과 비교했을 때의 많은 장점으로 인해 기체 센서로서 널리 사용된다[15]. 센서의 각 채널들은 기체의 종류에 따라 각각 다른 반응을 보인다. 센서의 측정(measurement)은 크게 안정화(stabilization), 노출(exposure), 배출(purge)의 세 단계로 구성되는데 각 단계에서의 반응을 모두 연결하여 한 개의 벡터로 기록한다[10].

컴퓨팅 시스템에서는 센서 어레이를 통해 입력된 기체 데이터의 종류를 분별하는 데에 여러 종류의 패턴 인식 기법을 사용한다. 기체를 분류하는 과정은 기체로부터 분류에 유용한 특징들을 추출하는 단계와 추출된 특징들을 이용하여 분류기를 설계하는 단계로 나눌 수 있다. 전자코 시스템에서 센서를 통해 얻은 데이터 샘플은 각 채널의 샘플링 지점(sampling point)마다의 측정값들로 구성된 고차원 데이터이다. 따라서 분류의 정확도를 높이면서 동시에 데이터 처리의 양을 효과적으로 줄이기 위해 다양한 종류의 차원 축소 방법을 이용할 수 있다. 대표적인 차원 축소 방법으로서 PCA(Principal Component Analysis)[16]는 데이터 샘플들의 분산을 가장 크게 만드는 저차원의 특징 공간에 데이터 샘플을 투영하여 분류를 수행한다. LDA(Linear Discriminant Analysis)[17]는 최적의 선형 분류 함수(linear discriminant function)를 찾는데 초점을 맞추는 방법으로서, 일반적으로 분류 문제에 대해 PCA보다 좋은 성능을 보인다. LDA는 학습(training) 데이터의 클래스 정보를 활용하여 클래스 내의 분포를 좁히고 클래스 간의 거리를 넓히는 특징 공간을 구성한다. [18]과 [19]에서는 PCA의 변형들 가운데 하나인 CC-PCA와 CC-CPCA를 각각 전자코 시스템에 사용하였으며, [10]은 LDA를 입력변수(input variable)들의 상관관계(correlation)가 큰 고차원 데이터에 적합하게 발전시킨 CLDA(Composite LDA)[20]를 기체 분류에 사용하였다.

이러한 방법들은 기체 샘플에 대해 안정화 단계에서부터 배출 단계에 이르기까지의 전체 측정값들을 한 개의 벡터로 표현하고 통계적인 분석 과정을 통해 특징 공간을 구성한다. 그런데, 센서 측정 과정에서 일부 데이터가 손실되거나 노이즈가 발생하는 등의 문제가 발생할 경우, 전체 측정값들을 한 개의 벡터로 한꺼번에 사용할 경우, 데이터 자체의 오류가 분류 결과에 그대로 영향을 끼칠 수 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 전체 측정 구간에서 추출한 전역특징 뿐만 아니라 측정의 각 단계(안정화, 노출, 배출)에 해당하는 지역 구간별로 지역특징을 추출하였다. 그런 다음, 추출된 지역특징들은 판별 분석에 기반을 둔 특징 선택 방법[21]을 이용하여 각 특징들이 기체를 분류하는데 얼마나 유용한지를 측정하여 그 측정 결과를 바탕으로 분류에 유용한 정보가 많은 특징들만을 선별하여 전역특징과 함께 결합 특징을 구성하였다.

제안한 방법은 특징 선택 과정을 통해 여러 구간에서 분별력 정보를 유지하고 있는 특징들만을 선별하여 사용하기 때문에

데이터 측정 과정에서 문제가 발생하여 데이터의 손실이 있는 경우에도 강한 동작을 하는 장점이 있다. 여러 종류의 VOC(Volatile Organic Compounds)에 대한 실험 결과를 통해 제안한 결합 특징이 전역특징을 사용하는 것보다 더 좋은 분류 성능을 나타내는 것을 확인할 수 있었다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 전자코 시스템에 사용할 수 있는 분별력 있는 특징 추출(discriminant feature extraction) 방법을 설명한다. 3장에서는 판별 분석에 기반하여 지역 구간별로 추출한 특징들의 분별력을 측정하여 결합 특징을 구성하는 방법을 제안하고, 4장에서는 전자코 기체 분류 실험의 결과를 제시한다. 마지막으로 5장에서 결론을 맺는다.

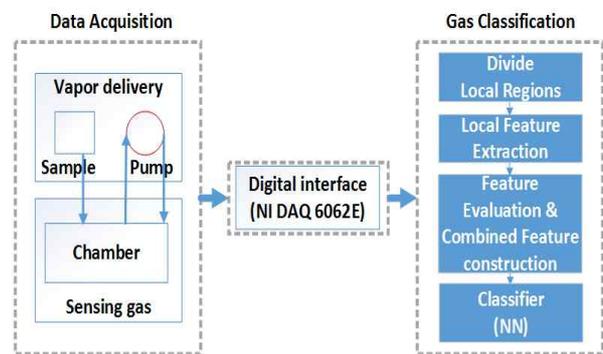


Fig. 1. Schematic Diagram of electronic nose system

II. Gas Classification Using Combined Features

1. Data Acquisition

Fig. 1은 본 논문에서 사용한 전자코 시스템의 전체 구성도(schematic diagram)이다. 기체의 측정은 [15]에서 사용한 마이크로머신(micromachined) 센서 어레이 칩을 사용하였다. 16의 채널로 구성된 센서 어레이는, 각 채널마다 맞물린 전극(interdigitated electrode), 마이크로히터(microheater), 머신 멤브레인(machined membrane)을 가진 CB(Carbon-Black) 고분자 복합물질 센서가 있다. Table 1에 16종류의 CB 고분자 복합물질들을 표시하였다. 각 고분자 복합물질 필름의 저항 변화는 화학 기체의 결합에 반응하여 관찰되는데 이를 0.1초마다 기록하여 총 200초 동안 측정하였다. 전체 측정구간은 안정화(30초)와 노출(60초), 배출(110초)의 세 개의 구간으로 나눌 수 있는데, 먼저 센서 어레이가 챔버(chamber)에 놓인 후에 저항 신호가 안정화되면, 시스템의 흐름 제어기(flow control)가 기체를 60초 동안 유입시키고, 나머지 110초 동안 잔여 기체를 외부로 배출시킨다[22]. 측정된 데이터는 디지털 인터페이스(DAQ6062E)를 통해 PC에 저장된다. 전압분배기

(voltage-divider)는 -10V에서 10V까지 동작하고 16 증폭기들의 이득(gain)은 최대 DAQ 해상도에 대해 10으로 설정하였다[15].

2. Feature Extraction Based on Discriminant Analysis

센서 어레이를 통해 얻은 기체 데이터로부터 대표적인 분류를 위한 차원 축소 방법 가운데 하나인 LDA를 이용하여 분류에 적합한 특징을 추출하였다. C 개의 클래스를 가진 N 개의 n 차원의 학습 데이터 샘플 $\mathbf{x}_k \in \mathbb{R}^{n \times 1}$, ($k = 1, \dots, N$)가 주어졌을 때, LDA는 같은 클래스에 속한 샘플들은 모으면서 동시에 클래스의 평균들 사이의 거리는 멀어지게 하는 특징 공간을 구성

Table 1. CB Polymer Composites in the Sensor Array[15]

Ch.	Polymer
1	Poly(Methyl methacrylate)
2	Polyvinylpyrrolidone
3	Poly(vinyl acetate)
4	Poly(ethylene oxide)
5	Polycaprolactone
6	Poly(4-methylstyrene)
7	Poly(styrene-co-methyl methacrylate)
8	Poly(ethylene-co-vinylacetate)
9	Poly(bisphenol A carbonate)
10	Poly(4-vinyl pyridine)
11	Poly(vinyl butyral)-co-vinyl alcphol-co-vinyl acetate
12	Poly(vinyl stearate)
13	Ethyl cellulose
14	Polystyrene-black-polyisoprene-black-polystyrene
15	Hydroxypropyl cellulose
16	Cellulose acetate

한다. 이를 위해, 먼저 클래스 내 산란 행렬(within-class scatter matrix) S_W 와 클래스 간 산란 행렬(between-class scatter matrix) S_B 를 다음과 같이 정의하였다[17].

$$S_B = \frac{1}{N} \sum_{i=1}^C N_i (\mu_i - \mu) (\mu_i - \mu)^T \quad (1)$$

$$S_W = \sum_{i=1}^C \sum_{\mathbf{x}_k \in c_i} (\mathbf{x}_k - \mu_i) (\mathbf{x}_k - \mu_i)^T$$

여기서 \mathbf{x}_k 는 클래스 c_i 에 속한 m 번째 샘플을 의미하며, N 은 전체 샘플의 수, C 는 클래스의 개수, 그리고 μ 와 μ_i 는 전체 샘플의 평균(mean)과 클래스 c_i 에 속한 샘플들의 평균이다. S_W 와

S_B 로부터 LDA는 다음과 같이 정의된 목적함수를 만족하는 투영벡터 (\mathbf{w}_l^L , $l = 1, \dots, n'$)들로 구성된 투영행렬(projection matrix) $W_{LDA} = [\mathbf{w}_1^L, \dots, \mathbf{w}_n^L]$ 을 구한다.

$$W_{LDA} = \operatorname{argmax}_W \frac{|W^T S_B W|}{|W^T S_W W|} \quad (2)$$

식(2)의 \mathbf{w}_l^L 들은 $S_B \mathbf{w}_l^L = \lambda_l S_W \mathbf{w}_l^L$ 를 만족하며 이는 $s_w^{-1} s_b$ 의 고유벡터(eigenvector)를 계산하여 얻을 수 있다[17]. 그런데, 기체 데이터(본 논문에서, $\mathbf{x}_k \in \mathbb{R}^{32000}$)와 같이 고차원 데이터의 경우에는 n 이 S_W 의 rank ($N - C$)보다 커지므로 S_W 가 항상 특이(singular)해지는 SSS(Small Sample Size)문제[17]가 발생한다. 이러한 문제를 해결하기 위한 방법으로서, PCA+LDA[23] 방법은 LDA를 수행하기에 앞서 PCA를 수행하여 데이터의 차원을 $m \leq N - C$ 로 줄인다. 전체 산란 행렬(total scatter matrix) S_T 를 다음과 같이 정의하면

$$S_T = \sum_{i=1}^C \sum_{\mathbf{x}_k \in c_i} (\mathbf{x}_k - \mu) (\mathbf{x}_k - \mu)^T, \quad (3)$$

PCA+LDA의 투영행렬은 $W_{PCA+LDA} = W_{LDA}^T W_{PCA}^T$ 가 된다. 여기서 $W_{PCA} = \operatorname{argmax}_W |W^T S_T W|$,

$$W_{LDA} = \operatorname{argmax}_W \frac{|W^T W_{PCA}^T S_B W_{PCA} W|}{|W^T W_{PCA}^T S_W W_{PCA} W|} \text{ 이다.}$$

$W_{PCA+LDA}$ 를 구성하고 있는 투영벡터 \mathbf{w}_i^{PL} 들 중에서 고유값(eigenvalue)이 큰 순서대로 $n' \leq C - 1$ 를 선택하면, 기체 데이터 샘플 \mathbf{x}_k 는 n' 개의 판별 특징(discriminant feature)들로 구성된 n' 차원의 특징벡터(\mathbf{y}_k)로 다음과 같이 표현된다.

$$\mathbf{y}_k = W_{PCA+LDA}^T (\mathbf{x}_k - \mu) = [y_{k1}, y_{k2}, \dots, y_{kn'}]^T \quad (4)$$

3. Feature Evaluation Based on Discriminant Analysis

제안한 방법에서는 먼저 전체 측정값을 모두 담고 있는 전역 샘플($\mathbf{x}_k^G \in \mathbb{R}^{32000 \times 1}$)로부터 안정화, 노출, 배출 구간에만 해당하는 측정값을 각각 분리하여 지역 샘플 $\mathbf{x}_k^{\text{st}} \in \mathbb{R}^{4800 \times 1}$, $\mathbf{x}_k^{\text{ex}} \in \mathbb{R}^{9600 \times 1}$, $\mathbf{x}_k^{\text{pu}} \in \mathbb{R}^{17600 \times 1}$ 를 구성하였다. 그런 다음, 식 (4)을 이용하여 전역특징 $\mathbf{y}_k^G = [y_{k1}^G, \dots, y_{kn'}^G]^T$ 과 구간 별 지역 특징 $\mathbf{y}_k^{\text{st}} = [y_{k1}^{\text{st}}, \dots, y_{kn'}^{\text{st}}]^T$, $\mathbf{y}_k^{\text{ex}} = [y_{k1}^{\text{ex}}, \dots, y_{kn'}^{\text{ex}}]^T$, $\mathbf{y}_k^{\text{pu}} = [y_{k1}^{\text{pu}}, \dots, y_{kn'}^{\text{pu}}]^T$ ($n' = C - 1$)을 구하고, 지역특징들로 3 ($C - 1$)개의 원소를 갖는 기본특징 벡터 $\mathbf{Y}_k^{\text{bol}} = [Y_{k,1}, \dots, Y_{k,3(C-1)}]^T (= [y_{k,1}^{\text{ex}}, \dots, y_{k,C-1}^{\text{ex}}, y_{k,1}^{\text{pu}}, \dots, y_{k,C-1}^{\text{pu}}]^T)$ 을 구성하였다.

각각의 기본특징들이 패턴을 분류하는데 얼마나 유용한지를 측정하기 위해 대표적인 특징 선택 방법인 Relief를[21] 사용하였다. Relief 방법은 각 특징들의 값이 서로 가까이 있는 샘플들을 잘 구분할 수 있는지를 정량적으로 측정하여 유용한 특징과 그렇지 않은 특징을 구분한다[24].

먼저, 각 기본특징 벡터와 동일한 크기의 참조벡터 $\mathbf{A} = \{A_1, A_2, \dots, A_{3(C-1)}\}^T$ 를 정의하고 0으로 초기화 한다. 그런 다음, N 개의 각 학습 데이터 샘플(\mathbf{x}_k)에 해당하는 기본특징 벡터(\mathbf{Y}_k^{Pool})를 가지고 기본특징들($Y_{k,i}^{Pool}, i = 1, \dots, 3(C-1)$)에 대한 유용성을 평가하여 그 결과를 참조벡터에 기록한다. \mathbf{Y}_k^{Pool} 에 대해, \mathbf{Y}_k^{Pool} 와 같은 클래스에 속한 학습 데이터 샘플들과 다른 클래스에 속한 샘플들 중에서 기본특징 공간에서 거리가 가장 가까운 v 개의 샘플을 각각 ‘S’($\mathbf{S}_r = [S_{r,1}, \dots, S_{r,3(C-1)}]^T$)과 ‘D’

\mathbf{Y}_k^{Pool} 의 j 번째 기본특징 $Y_{k,j}$ 이 분별력이 있는 특징이라면 $Y_{k,j}$ 는 $S_{r,j}$ 와는 비슷한 값을 갖고 반면 $D_{r,j}$ 와는 확연히 구분이 되어야 한다. 따라서 $|Y_{k,j} - S_{r,j}|$ 와 $|Y_{k,j} - D_{r,j}|$ 으로부터 식 (5)에 의해 계산된 값으로 참조벡터의 j 번째 원소 A_j 의 값을 업데이트한다. A_j 값을 업데이트 하는 과정은 다음과 같다.

- 1) 데이터 샘플 \mathbf{x}_k 로부터 \mathbf{Y}_k^{Pool} 를 구한다.
- 2) 기본특징 공간에서 v 개의 \mathbf{S}_r 를 찾는다.
- 3) \mathbf{Y}_k^{Pool} 의 클래스 $c_{\mathbf{Y}_k^{Pool}}$ 와 다른 클래스들 c_l ($l = 1, \dots, C-1$)에서 v 개의 $\mathbf{D}_r(c_l)$ 들을 찾는다.
- 4) $A_j = A_j - \sum_{r=1}^v |Y_{k,j} - S_{r,j}| / (N \cdot s) + \sum_{c_l \neq c_{\mathbf{Y}_k^{Pool}}} \left[\frac{1 - P(c_l)}{1 - P(c_{\mathbf{Y}_k^{Pool}})} \sum_{r=1}^v |Y_{k,j} - D_{r,j}(c_l)| \right] / (N \cdot s)$ (5)

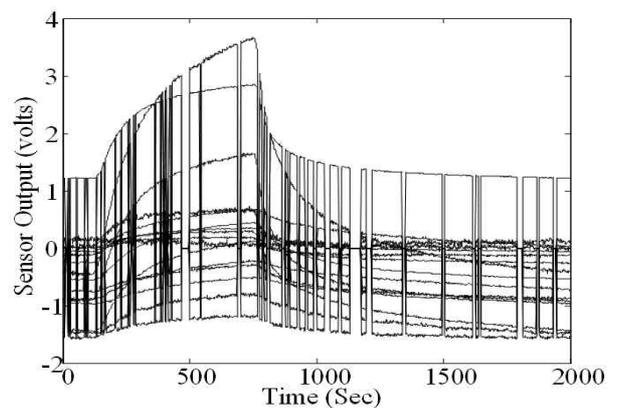
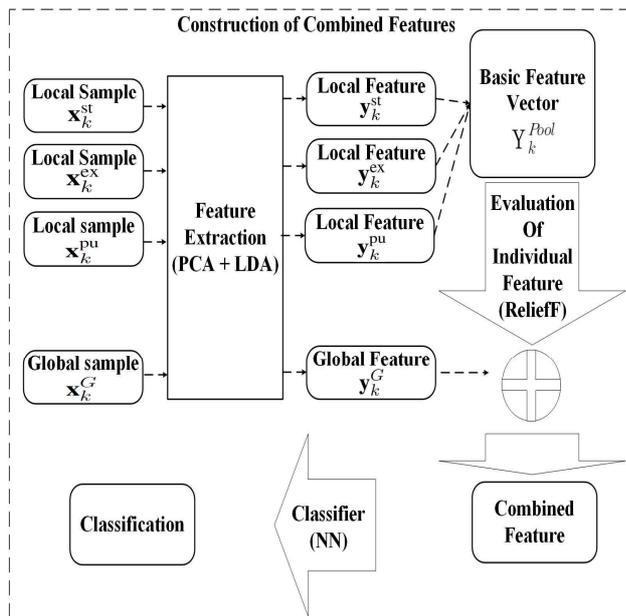
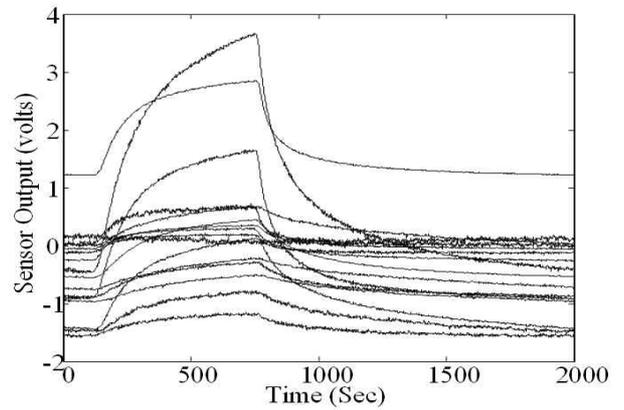
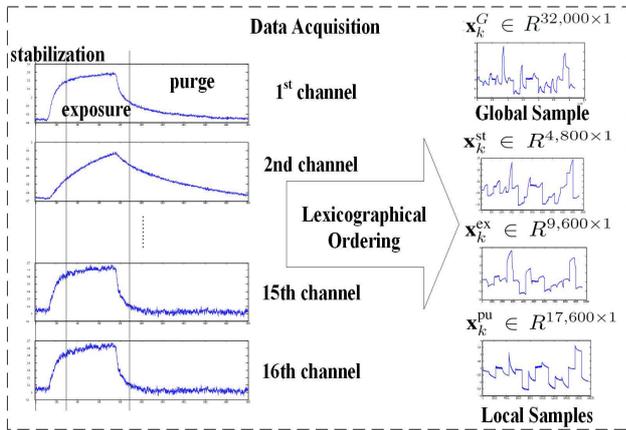


Fig. 2. Overall Procedure of Gas Classification Using Combined Features

Fig. 3. Electronic nose data: (a) Lossless data. (b) 20% data loss.

($\mathbf{D}_r = [D_{r,1}, \dots, D_{r,3(C-1)}]^T$) ($r = 1, \dots, v$)로 정의한다[21]. 만약

여기에서 $P(c_l)$ 는 클래스 c_l 의 사전 확률(prior probability)이다. 위의 과정을 N 개의 샘플 \mathbf{x}_k ($k = 1, 2, \dots, N$)에 대해 반복한 후, A_j 의 값이 큰 v 개의 특징들($Y_t, t = 1, \dots, v$)을 가지고 전역특

징과 함께 기체 분류에 사용 될 최종 결합 특징 벡터 ($\mathbf{y}^{CF} = [y_{k1}^G, \dots, y_{kn}^G, Y_{k1}, \dots, Y_{kn}]^T$)를 구성하였다. 제안한 방법은 여러 구간에서의 지역특징들 중에서 분별력이 높은 특징들만 선별하여 결합 특징을 구성함으로써, 각 구간별로 불필요한 정보들은 제거하고 유용한 특징들만 분류기의 입력으로 사용한다. 이로 인해 기체 데이터를 측정하는 과정에서 일부 손실이 발생한 경우에도 특징 선택 과정에서 데이터 손실의 영향을 받은 특징들이 제거됨으로써 분류 성능이 감소하는 것을 방지할 수 있다. 제안한 방법의 전체 흐름은 Fig. 2에 제시하였다.

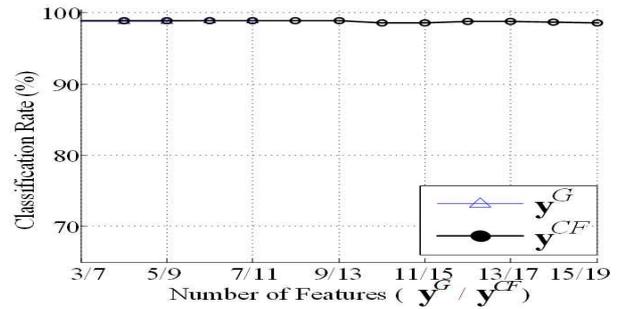
III. Experimental Results

1. Experimental Condition

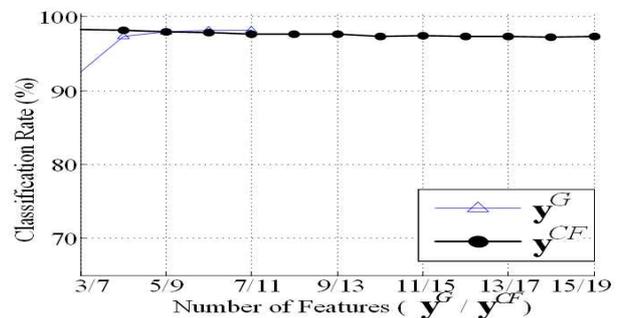
제안한 방법이 전자코 시스템에서 기체를 효과적으로 잘 분류할 수 있는지 확인하기 위해 총 8개 종류의 기체에 대한 VOC 측정 데이터에 대한 분류 실험을 해 보았다. 8개 기체의 종류는 acetone, benzene, cyclo-hexane, ethanol, heptane, methanol, propanol, toluene이다[15]. 기체의 종류별로 20개의 샘플을 채취하여 총 160개의 샘플을 실험에 사용하였다. 각각의 샘플은 한 채널당 10Hz의 샘플링 비율로 200초 동안 측정된 2000개 시점에 대한 측정치로 구성되어 있으며, 16개 채널의 측정치를 2000x16 크기의 행렬 형태로 저장한 후 렉시코 그래픽 정렬 연산자(lexicographic ordering operator)를 이용하여 32000차원의 벡터로 변환하였다[25]. 제안한 방법이 측정 데이터에 손실이 발생하였을 때 얼마나 효과적인지를 확인하기 위해, 무손실 데이터 샘플과, 손실 양이 전체 측정 지점의 10%, 20%, 30%에 해당하는 손실 데이터 샘플(Fig. 3)에 대해 성능을 분석해 보았다.

총 160개의 데이터 샘플에 대한 분류 성능은 8-fold 교차 검증 방법(cross validation strategy)[26,27]을 이용하여 평가하였다. 먼저 160개의 샘플을 시드(seed) 값 기반의 난수 발생기(random number generator)를 이용하여 섞은 후, 같은 개수의 샘플이 포함되도록 8개의 fold로 나누고 그룹마다 1개의 샘플씩 테스트 샘플로 사용하고 (즉 총16개의 샘플), 나머지 140개의 샘플을 학습(training) 데이터로 사용하여 분류 율을 구하였다. 같은 방식으로 각 fold의 모든 데이터 샘플이 한 번씩 테스트 샘플로 사용되도록 10번의 실험을 수행하여 분류 율의 평균값을 구하였다. 본 논문에서는 8-fold 교차 검증 방법의 통계적인 신뢰도를 높이기 위해, 이와 같은 과정을 서로 다른 8개의 난수 시드 숫자(random seed number)를 이용하여 8번 반복하여 얻은 결과의 평균값을 분류 성능의 지표로 사용하였다. 모든 데이터 샘플은 학습 데이터의 평균과 표준편차로 정규화 하여 사용하였다. PCA+LDA를 이용하여 판별 특징을 추출할 때, PCA를 이용하여 얼마의

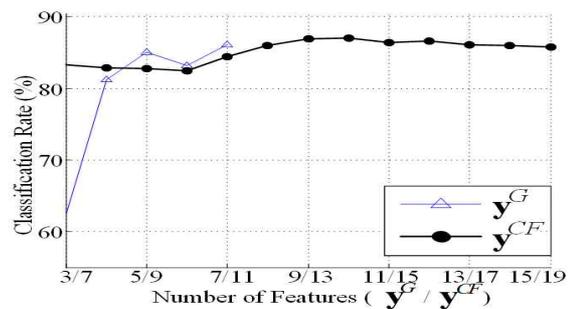
차원(m)으로 축소하는지는 판별 특징의 분류 성능에 큰 영향을 미친다. 본 실험에서는 여러 종류의 m 값에 대한 성능 평가를 바탕으로 S_T 의 전체 고유값의 99%를 차지하도록 m (=105)을 정하였다. 분류기로는 NN 규칙(Nearest Neighbor rule)을 사용하였으며, 두 샘플 사이의 거리 계산은 l_2 노름(norm)을 사용하였다[25,28].



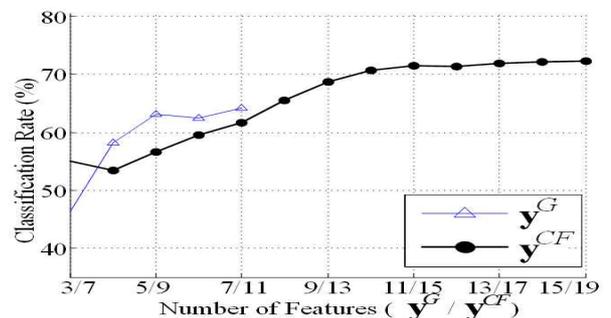
(a)



(b)



(c)



(d)

Fig. 4. Classification Rates of Global Feature and

Combined Feature for the Degree of Data Loss: (a) lossless. (b) 10% data loss. (c) 20% data loss. (d) 30% data loss.

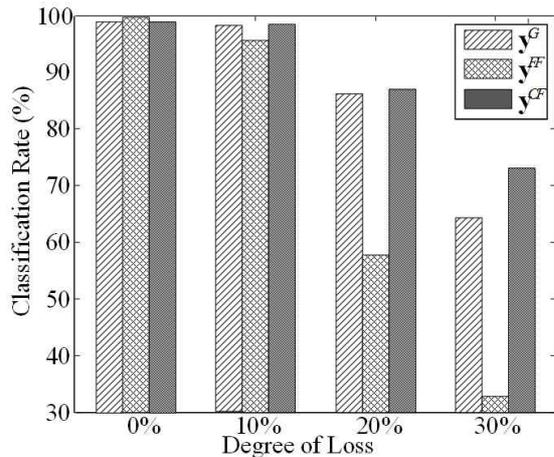


Fig. 5. Comparison of Classification Rates Between the Proposed Method and Other Methods

2. Classification Rates

데이터 측정 과정에서 발생한 문제로 인해 전체 데이터 측정치의 10%, 20%, 30%의 손실이 생겼을 경우를 가정하여, 전체 측정값을 갖는 샘플 (\mathbf{x}_k^G)로부터 추출한 전역특징(\mathbf{y}^G)과 세 종류의 선별된 지역특징들(\mathbf{y}_k^{st} , \mathbf{y}_k^{ex} , \mathbf{y}_k^{pu})을 함께 사용한 결합 특징(\mathbf{y}^{CF})의 성능을 평가해 보았다. 8개의 클래스를 갖는 문제에서 PCA+LDA는 최대 7개의 특징을 추출할 수 있으므로 \mathbf{y}^G 의 경우 특징벡터의 최대 차원은 7이며, \mathbf{y}^{CF} 의 경우는 \mathbf{y}_k^{st} , \mathbf{y}_k^{ex} , \mathbf{y}_k^{pu} 에서의 지역특징들이 모두 결합되면 최대 28차원이 된다.

Fig. 4는 손실 데이터에 대해 \mathbf{y}^G 와 \mathbf{y}^{CF} 의 특징 차원 수에 따른 분류율을 나타내고 있다. 그림 4(a)-(b)와 같이 어느 정도의 손실의 양까지는 \mathbf{y}^G 와 \mathbf{y}^{CF} 모두 높은 분류율을 나타내는 것을 볼 수 있다. 그러나 데이터 손실의 양이 20%에서 30%로 점점 커짐에 따라 결합 특징 \mathbf{y}^{CF} 이 \mathbf{y}^G 보다 높은 분류 성능을 보였다. 한편, \mathbf{y}^{CF} 의 차원이 증가함에 따라 \mathbf{y}^{CF} 의 분류율도 함께 증가하다가 어느 시점부터는 오히려 분류율이 감소하는 것을 볼 수 있다. 이는 모든 지역특징들이 분류에 도움이 되는 것이 아니며, 특징 선택 방법을 통해 분류에 방해가 되는 특징들이 효과적으로 제거될 수 있음을 보여준다.

Fig. 5에서는 제안한 결합특징(\mathbf{y}^{CF})과 전역특징(\mathbf{y}^G), FF(Feature Feedback)[25]에 의한 특징(\mathbf{y}^{FF})에 대한 분류 성능을 비교하였다. Fig. 5에서 볼 수 있듯이, 데이터 손실이 없을 때에는 \mathbf{y}^{FF} 이 \mathbf{y}^G 이나 \mathbf{y}^{CF} 보다 다소 높은 분류율을 보여 주었으나, 다양한 정도의 데이터 손실에서 제안한 방법이 다른 방법보다 더 안정적인 성능을 보이는 것을 볼 수 있다. 이는 제안

한 방법이 특징 선택과정을 통해 분별력이 좋은 여러 종류의 지역특징들을 선별적으로 전역특징과 함께 분류기에 사용하였기 때문으로 보인다.

IV. Conclusions

전자코 시스템에서 센서를 이용한 기체 데이터의 측정은 그 과정에 따라 여러 단계로 구분 지을 수 있다. 각 단계에서의 센서의 반응은 각각 다른 정보를 가지고 있어 이로부터 분류를 위한 다양한 특징들을 추출할 수 있다. 본 논문에서는 안정화, 노출, 배출 단계로 구분된 구간에서 세 종류의 지역특징들을 추출하였다. 개별 지역특징들에 대해 관별 분석에 기반을 두어 특징의 분별력을 평가하여, 분별력 정보 양이 많은 특징들만 선별하여 전역특징과 함께 분류기에 사용할 결합 특징을 구성하였다. 16개의 채널로 구성된 센서 어레이를 이용하여 측정된 8종류의 기체 데이터에 대한 실험에서, 제안한 방법은 좋은 분류 성능을 보여 주었으며, 특히 센싱 과정에서 일부 데이터 손실이 발생한 경우에 전역특징, 또는 기체 분류를 위한 다른 특징 추출 방법보다 더 우수한 분류 성능을 보여 주었다.

향후, 보다 다양한 종류의 복잡한 데이터에 대한 실험과 함께 여러 종류의 특징들을 효과적으로 결합하는 연구도 진행할 예정이다.

REFERENCES

- [1] S.-W. Chiu, and K.-T. Tang, "Towards a Chemiresistive Sensor-integrated Electronic Nose: A Review," *Sensors*, Vol. 13, No. 10, pp. 14214-14247, 2013.
- [2] J. Zhou, T. Feng, and R. Ye, "Differentiation of Eight Commercial Mushrooms by Electronic Nose and Gas Chromatography-Mass Spectrometry," *Journal of Sensors*, Vol. 2015, 2015.
- [3] M. Bougrini, K. Tahri, Z. Haddi, T. Saidi, N. El Bari, and B. Bouchikhi, "Detection of Adulteration in Argan Oil by Using an Electronic Nose and a Voltammetric Electronic Tongue," *Journal of Sensors*, Vol. 2014, August, 2014.
- [4] A.D. Wilson and M. Baietto, "Applications and Advances in Electronic-nose Technologies," *Sensors*, Vol. 9, No. 7, pp. 5099-5148, 2009.
- [5] S.-I. Choi and G.-M. Jeong, "A Discriminantant

- Distance Based Composite Vector Selection Method for Odor Classification,” *Sensors*, Vol. 14, No. 4, pp. 6938–6951, 2014.
- [6] J. Lerchner, D. Caspary, and G. Wolf, “Calorimetric Detection of Volatile Organic Compounds,” *Sensors and Actuators B: Chemical*, Vol. 70, No. 1, pp. 57–66, November 2000.
- [7] M. Farre, L. Kantiani, M. Petrovic, S. Perez, and D. Barcelo, “Achievements and Future Trends in the Analysis of Emerging Organic Contaminants in Environmental Samples by Mass Spectrometry and Bioanalytical Techniques,” *Journal of chromatography A*, Vol. 1259, pp. 86–99, October 2012.
- [8] Y.-H. Kim, and K.-H. Kim, “Ultimate Detectability of Volatile Organic Compounds: How Much Further Can We Reduce Their Ambient Air Sample Volumes for Analysis?,” *Analytical Chemistry*, Vol. 84, No. 19, pp. 8284–8293, August 2012.
- [9] J. Nicolas, A.-C. Romain, V. Wiertz, J. Maternova, and P. Andre, “Using the Classification Model of an Electronic Nose to Assign Unknown Malodours to Environmental Sources and to Monitor Them Continuously,” *Sensors and Actuators B: Chemical*, Vol. 69, No. 3, pp. 366–371, October 2000.
- [10] S.-I. Choi, G.-M. Jeong, and C. Kim, “Classification of Odorants in the Vapor Phase Using Composite Features for a Portable E-nose System,” *Sensors*, Vol. 12, No. 12, pp. 16182–16193, 2012.
- [11] C. Di Natale, A. Macagnano, E. Martinelli et al., “Lung Cancer Identification by the Analysis of Breath by Means of an Array of Non-selective Gas Sensors,” *Biosensors and Bioelectronics*, Vol. 18, No. 10, pp. 1209–1218, September 2003.
- [12] K. Arshak, E. Moore, G.M. Lyons, J. Harris, and S. Clifford, “A Review of Gas Sensors Employed in Electronic Nose Applications,” *Sensor Review*, Vol. 24, No. 2, pp. 181–198, 2004.
- [13] K.J. Albert, N.S. Lewis, C.L. Schauer, G.A. Sotzing, S.E. Stitzel, T.P. Vaid, and D.R. Walt, “Cross-reactive Chemical Sensor Arrays,” *Chemical Reviews*, Vol. 100, No. 7, pp. 2595–2626, 2000.
- [14] A. Šetkus, A. Olekas, D. Senulienė, M. Falasconi, M. Pardo, and G. Sberveglieri, “Analysis of the Dynamic Features of Metal Oxide Sensors in Response to SPME Fiber Gas Release,” *Sensors and Actuators B: Chemical*, Vol. 146, No. 2, pp. 539–544, April 2010.
- [15] Y.S. Yang, S.-C. Ha, and Y.S. Kim, “A Matched-profile Method for Simple and Robust Vapor Recognition in Electronic Nose (E-nose) System,” *Sensors and Actuators B: Chemical*, Vol. 106, No. 1, pp. 263–270, April 2005.
- [16] M. Turk, and A. Pentland, “Eigenfaces for Recognition,” *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71–86, 1991.
- [17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, NY, USA, 2nd edition, 1990.
- [18] T. Artursson, T. Eklov, I. Lundstrom, P. Martensson, M. Sjostrom, and M. Holmberg, “Drift Correction for Gas Sensors Using Multivariate Methods,” *Journal of Chemometrics*, Vol. 14, No. 5–6, pp. 711–723, 2000.
- [19] 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, pp. 460–265, April 2010.
- [20] C. Kim and C.-H. Choi, “A discriminant analysis using composite features for classification problems,” *Pattern Recognition*, Vol. 40, No. 7, pp. 2118–2125, November 2007.
- [21] M. Robnik-Šikonja, and I. Kononenko, “Theoretical and Empirical Analysis of Relief and RRelief,” *Machine Learning*, Vol. 53, No. 1–2, pp. 23–69, October 2003.
- [22] S.-C. Ha, Y.S. Kim, Y. Yang, Y.J. Kim, S.-M. Cho, H. Yang, and Y.T. Kim, “Integrated and Microheater Embedded Gas Sensor Array Based on the Polymer Composites Dispensed in Micromachined Wells,” *Sensors and Actuators B: Chemical*, Vol. 105, No. 2, pp. 549–555, March 2005.
- [23] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp.711–720, July 1997.
- [24] S.-I. Choi, “Construction of Composite Feature Vector Based on Discriminant Analysis for Face Recognition,” *Journal of Korea Multimedia Society*,

Vol. 18, No. 7, pp. 834-842, July 2015.

- [25] S.-I. Choi, S.-H. Kim, Y. Yang, and G.-M Jeong, "Data Refinement and Channel Selection for a Portable E-nose System by the Use of Feature Feedback," *Sensors*, Vol. 10, No. 11, pp. 10387-10400, 2010.
- [26] L. Liu, and M.T. Zsu, *Encyclopedia of Database Systems*, Springer Publishing Company, Springer: New York, NY, USA, 2009.
- [27] G.Y. Heo, H. Choi, and J.S. Youn, "Supervised Rank Normalization with Training Sample Selection," *Journal of the Korea Society of Computer and Information*, Vol. 20, No. 1, pp. 21-28, January 2015.
- [28] S.-I. Choi, "Hybrid Pattern Recognition Using a Combination of Different Features," *Journal of the Korea Society of Computer and Information*, Vol. 20, No. 1, pp. 21-28, Vol. 20, No. 11, 2015.

Authors



Sang-Il Choi received the B.S. degree in the division of electronic engineering from Sogang University, Seoul, Korea, in 2005 and the Ph.D. degree from the School of Electrical Engineering and Computer Science, Seoul National University, Seoul, in 2010.

He was a Postdoctoral Researcher in the BK21 Information Technology, Seoul National University, in 2010 and in the Institute for Robotics and Intelligent Systems of Computer Science Department, University of Southern California, Los Angeles, until August of 2011. He is currently an Associate Professor with the Department of Computer Science and Engineering, Dankook University, Gyeonggi-do, Korea. His research interests include pattern recognition, feature extraction and selection, machine learning, computer vision, and their applications.