<http://www.diva-portal.org>

Preprint

This is the submitted version of a paper published in *Sensors and actuators. B, Chemical*.

Citation for the original published paper (version of record):

Fan, H., Hernandez Bennetts, V., Schaffernicht, E., Lilienthal, A. (2018)
A cluster analysis approach based on exploiting density peaks for gas discrimination
with electronic noses in open environments
Sensors and actuators. B, Chemical, 259: 183-203
<https://doi.org/10.1016/j.snb.2017.10.063>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-63468>

©2018. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

A Cluster Analysis Approach Based on Exploiting Density Peaks for Gas Discrimination with Electronic Noses in Open Environments

Han Fan, Victor Hernandez Bennetts, Erik Schaffernicht and Achim J. Lilienthal

AASS Research Centre. Örebro University. SE-70182, Örebro, Sweden

Abstract

Gas discrimination in open and uncontrolled environments based on smart low-cost electro-chemical sensor arrays (e-noses) is of great interest in several applications, such as exploration of hazardous areas, environmental monitoring, and industrial surveillance. Gas discrimination for e-noses is usually based on supervised pattern recognition techniques. However, the difficulty and high cost of obtaining extensive and representative labeled training data limits the applicability of supervised learning. Thus, to deal with the lack of information regarding target substances and unknown interferences, unsupervised gas discrimination is an advantageous solution. In this work, we present a cluster-based approach that can infer the number of different chemical compounds, and provide a probabilistic representation of the class labels for the acquired measurements in a given environment. Our approach is validated with the samples collected in indoor and outdoor environments using a mobile robot equipped with an array of commercial metal oxide sensors. Additional validation is carried out using a multi-compound data set collected with stationary sensor arrays inside a wind tunnel under various airflow conditions. The results show that accurate class separation can be achieved with a low sensitivity to the selection of the only free parameter, namely the neighborhood size, which is used for density estimation in the clustering process.

Keywords: gas discrimination; environmental monitoring; metal oxide sensors; cluster analysis; unsupervised learning.

Introduction

Gas discrimination using low-cost electronic noses (e-noses) is of great interest in many applications outside laboratory conditions, such as detecting hazardous gases, monitoring air pollution in urban areas, emissions from sewage facilities and animal production facilities [1-3]. In some cases, these applications entail continuous data collection in open and complex environments. One prominent example is mobile robots equipped with e-noses conducting tasks like exploration of hazardous areas and leak detection [3-5].

Traditionally, gas discrimination with e-noses has been carried out in chambers, where environmental conditions, such as humidity, temperature, airflow, and gas exposure patterns are tightly controlled. Such conditions enable the well established three-phase sampling strategy to be performed [4]. In a three phase-sampling process, the sensors are firstly exposed to a reference gas (e.g. clean air) to set a known baseline response level for the sensor array. Then, the sensors interact with injected gas samples for a period of time until a steady response state is reached. The sampling process concludes as the sensors recover to their baseline level when the gas sample is flushed away. For example, in Fig. 1a, the sensor responses show a clear three-phase profile, which is acquired with the sensors inside in a chamber after the sensors have been exposed to the gas sample for a considerable amount of time. Up to date, many studies on gas discrimination have achieved great success under the three-phase sampling strategy [6-8]. Contrary to laboratory

conditions, in uncontrolled environments, e-noses are often directly exposed to dynamic conditions, such as varying temperature, humidity, and airflow patterns. These complex ambient conditions cause fluctuating gas concentration levels, and the sensor responses show intermittent and transient behavior instead of well defined three-phase patterns. Fig. 1b is an example, in which the measurements are taken by sensors mounted on a mobile robot exploring in a large room with an e-nose as a sensing payload. An obvious difference between Fig. 1b and Fig. 1a is that, the sensor responses in Fig. 1b never reached a steady state, which prevents the use of a three-phase sampling strategy [1].

Figure 1: Responses patterns acquired with MOX sensor arrays with different sampling processes. Each color represents the response of a single sensor. (a) An example of the three phase sampling process. The number in the figure indicate the stages of the sensor responses, namely 0-baseline response, 1-rising edge, 2-steady state, 3-recovery edge. The shaded area denotes the period of time during which the sensors are exposed to the chemical analyte; (b) Response pattern acquired in an uncontrolled environment. Both images adapted from [4].

Uncontrolled environments present a number of challenges to perform the three-phase sampling strategy. Usually gas discrimination is implemented by training a supervised classifier with steady signals, which are, however, rarely observed in open environments [4,5]. Although several works have been devoted to address this issue by considering the transient phase of the signal [4,8,9], it requires additional data processing to detect the rising and decaying edges in the instantaneous sensor responses. This process often leads to discarding valuable information since other sensor response phases are not considered. Also, preparing extensive and representative training data brings additional difficulties since obtaining and labeling samples are expensive and time consuming. More importantly, it is hard to recreate all possible situations in application scenarios, such as varying environmental conditions, and hardly feasible to acquire data that significantly represents all relevant chemical compounds. Aside from the known target analytes, unknown interferences, whose response patterns have not been included in the training set, might be present in the application environments. In practice, the resulting classification system is likely to fail when faced with new factors that were not taken into account in the training phase.

To overcome the aforementioned issues to some extent, an alternative solution is to perform clustering analysis to avoid relying on labeled training data only. As an unsupervised learning technique, clustering seeks to distinguish classes according to their relationship in the response pattern provided by the e-nose. However, clustering e-nose data in open environments is not trivial. One of the difficulties is that the inner pattern and structure of the e-nose data are rather complex. Take Fig. 2a for example, which shows the feature space of two typical data sets collected by a set of commercial MOX sensors in an uncontrolled environment. One

can notice the following: (1) observations from different classes overlap at low concentration levels; (2) the data are distributed in clusters of irregular shapes and measurement density; (3) the data set is not balanced with respect to the concentration level, as shown in the histograms in Fig. 2b). High-concentration data are sparse while most of the observations lie in low concentration regions. These high density, low-concentration observations correspond to the baseline responses of the sensors, when the array was exposed to clean air only. The above characteristics can pose obstacles for conventional clustering methods. For instance, K-means is not reliable to detect non-spherical clusters. Some density-based algorithms may declare the cluster centers at densely populated regions, which, in the case of gas discrimination, happens to be dominated by overlapping low-concentration measurements. In this way, the cluster centers are very likely to be not representative enough for corresponding classes.

To address these challenges, we present the *KmP* algorithm, which is a novel cluster-based approach tailored for gas discrimination in uncontrolled environments, where unknown interferents are likely to be present (Section 3). The *KmP* algorithm first automatically infers the number of chemical compounds from the data (*K*-learning phase). Further, it learns a classification threshold that separates the chemical substances from baseline responses (*m*-learning phase), and conducts cluster analysis accordingly. We suggest to combine the results of the cluster analysis and information in gas concentration levels together to estimate the class posterior (*P*-learning phase). We validate our approach with evaluation tests on a variety of open-environment data sets (Section 4). The evaluation results demonstrate the feasibility and parameter robustness of the *KmP* algorithm, as well as its superior performance compared to other classical clustering techniques (Section 5).

(a)

(b)

Figure 2: The feature space and the gas concentration histograms of measurements of propanol (green) and ethanol (red) collected in an open environment with three E2V manufactured metal oxide sensors (MOX) sensors (MICS-2710, MICS-5121, MICS-5521). **(a)**: The feature space plot. Each data point is an instantaneous sensor response, and the color shades denote the normalized concentration level of the measurements: **(b)**: Concentration levels shown in the histograms were obtained by a photo ionization detector (PID). The nested figures are the zoomed-in histograms with the ranges of high concentrations. Both data of propanol and ethanol are unbalanced with respect to the gas concentration.

Related Work

Gas Discrimination in Uncontrolled Environments

Several solutions have been proposed to the problem of gas discrimination in uncontrolled environments. Vergara *et al.* performed gas discrimination using inhibitory Support Vector Machines (SVM) [10]. They found that the classification result is influenced by the wind flow strength, and the distance between the sampling location and gas source. In their experiments, several chemical compounds were tested as the

target analyte under a series of different environmental conditions in a wind tunnel where e-noses composed of six commercial Figaro MOX sensors were placed at different distances from the gas outlet. An important observation made by the authors is that a robust gas discrimination system has to be trained with exhaustive environmental conditions, which in practice is not always feasible. As an extension work of [10], Fonollosa *et al.* [11] exposed a sensor array of eight Figaro MOX sensors to dynamic gas mixtures. In this experiment, the gas plumes were mixed naturally along the turbulent flow in the wind tunnel. The authors employed an inhibitory SVM as the tool to classify chemical components in dynamic turbulent mixtures, and the results showed that training data of high concentration alone is not enough to achieve an optimal classification performance. Thus, the authors concluded that the training of the classifier should consider the data at low concentrations as well. Capelli and co-authors developed a sophisticated e-nose system for outdoor odor monitoring in [2]. The sensor system was composed of two kinds of e-noses that are both equipped with 6 MOX sensors. One of the e-noses is developed mainly for laboratory use while the other performs better under variable meteorological conditions, and is more sensitive to diluted odors. The experiments included laboratory test and field tests. The field tests were carried out in an area around different industrial facilities, including two water treatment plants and an oil mill. The authors prepared the training data with the samples collected in two different days, and achieved a high classification success rate in identifying odors both in the samples coming from emissive plants as well laboratory tests. It is worth noticing that they found gas discrimination can be hardly performed for the diluted (i.e. low concentration) samples. This point is taken into account in the approach proposed in this paper.

Trincavelli and co-authors investigated the data collected with an Open Sampling System (OSS). The OSS used in their research is a mobile robot consisting of four Figaro TGS sensors. The robot patrolled to take measurements in a closed room [4,12] and an outdoor courtyard where gas sources were present [12]. The authors pointed out that in such data the steady state in the sensor response time series is rarely reached due to the fluctuating gas concentration levels. Therefore, the analysis should be based on the transient phase of the signal. However, considering the transient phase only neglects the information from other response phases. For example, stable low concentration measurements convey useful information to model the absence of gas in the environment. An approach that considers the whole signal was reported by Hernandez *et al.* [13]. Their approach considered the information from the gas concentration levels for class separation, which is implemented by taking the concentration levels of the observations into account to estimate class posteriors. The authors conducted experiments with a mobile robot equipped with a set of 6 E2V MOX sensors in a indoor robot arena, and a robot equipped with 4 Taguchi MOX sensors in an outdoor courtyard. The proposed approach not only achieved high classification success rates, but also addressed the problem of unbalanced data sets with respect to concentration levels. Such unbalanced data sets have very few high concentration measurements, while diluted measurements are present most of the time. This is a common issue with data collected in open environments. In a more recent work done by Eusebio *et al.* [14], the concentration level is considered as a factor affecting gas discrimination as well. The main purpose of this paper is to investigate the instrument capability to detect and classify several odors of interest. The authors describe an experimental procedure adopted to evaluate e-nose performance and the feasibility of supervised gas discrimination. From their work, we note that their classification success was achieved under the condition that the training set includes samples in a relatively wide range of concentration levels, while the test set contained samples with a smaller range of concentration level. Monroy *et al.* [3] conducted an experimental analysis focused on the role of motion speed of sampling platform in gas classification of OSS. Their experiment set-up was an indoor corridor with two different gas sources, and the sampling procedure is done by an robot carrying an e-nose (Figaro and Hanwei sensors) traversed the corridor back and forth at different speeds. The authors found that the classification accuracy is negatively correlated with the motion speed, and a supervised classifier trained with data collected in motion is more accurate than a classifier trained with data obtained when the sensing device is static. The works mentioned above show some typical challenges for gas discrimination in uncontrolled environments, such as diluted measurements and various environmental conditions. More importantly, supervised learning approaches, such the ones introduced above, have a key limitation: they can only be performed under a-priori assumption that the number and identities of the substances to discriminate are already known to the classification systems.

Unsupervised Gas Discrimination

In general, the applicability of supervised methods can be limited due to the need for sufficient training data. Compared to supervised learning, semi-supervised and unsupervised methods are rarely employed for gas discrimination. However, they can be useful alternatives. Recently, Jia *et al.* proposed a novel semi-supervised electronic nose learning technique to classify indoor pollution gases [15]. The experimental set-up of this work is rather ideal: in a temperature-humidity controlled chamber, the e-nose system, including 3 Figaro MOX sensors, was exposed to clear air for 2 minutes before the target gas was presented for 4 minutes, and then the sensor array was exposed to clean air for 9 minutes again to recover the baseline. Nevertheless, the highlight of this work is that the semi-supervised method can make the use of the unlabeled data to enhance training of classifiers. However, the training procedure still requires a certain amount of labeled samples. An approach that uses an unsupervised gas discrimination algorithm in complex scenarios is reported in [16]. Schleif *et al.* investigated a generative topographic mapping through time (GTM-TT) to model time-series for odor recognition. The classical GTM-TT modeling can be unsupervised, but it was further extended to supervised classification in [16]. This work is demonstrated to be effective for coping with high dimensional data, and the GTM-TT model is suitable for rapid odor classification problems in robotics. To our best knowledge, little attention has been given to unsupervised pattern recognition methods in gas discrimination problem under uncontrolled environmental conditions. A relatively close work is done by Brahim-Belhaouari and co-authors. They hybridized the classical K-means clustering and Gaussian mixture model with k -NN method for gas identification [17]. While the measurements of the gases was collected in a tightly controlled chamber by the commercial TGS and MHP gas sensors, their effort indicates that the classification system based on modified clustering algorithms can be very accurate and fast. In summary, unsupervised gas discrimination in open environments has seldom been systematically explored. Previous attempts either consider simplistic sampling scenarios, or partially depend on labeled training data, which means they share some of the limitations of supervised methods.

The KmP Algorithm

Before describing the KmP algorithm, we introduce the notation used in this section. A data set of gas concentration measurements is denoted as $\mathbf{X} = [\mathbf{r}_1^D, \mathbf{r}_2^D, \mathbf{r}_3^D, \dots, \mathbf{r}_N^D]$, containing N measurements of D dimensions. Each dimension of $\mathbf{r}^D$ corresponds to one sensor in the array. An unlabeled data set can be represented as $\mathbf{X} = \{\mathbf{A}, \mathbf{C}\}$, where $\mathbf{A}$ and $\mathbf{C}$ are subsets of $\mathbf{X}$. $\mathbf{C}$ represents measurements that are above a classification threshold m_D , which determines the concentration level at which gas discrimination can be performed. $\mathbf{A}$ is the subset of measurements that fall below the classification threshold m_D . It is thus assumed that the subset $\mathbf{A}$ corresponds to measurements taken under clean air conditions. Considering there are K^* target analytes in the subset $\mathbf{C}$, the probabilistic representation of the identity of the i^{th} measurement $\mathbf{r}_i^D$ in $\mathbf{X}$, is given by the class posterior $P(l|\mathbf{r}_i^D)$, where $l \in \{L_0, L_1, L_2, L_3, \dots, L_{K^*}\}$, indicates that the measurement belongs to one of the $K^* + 1$ substances, with L_0 corresponding to clean air.

The KmP algorithm is an integrated cluster analysis approach comprising three sequential steps, namely the K -learning phase, the m -learning phase, and the P -learning phase. The first two phases learn the number of clusters K^* and the classification threshold m_D respectively, while the P -learning phase determines the class posteriors for each measurement. As previously stated, m_D is the intrinsic parameter that separates $\mathbf{X}$ into subsets $\mathbf{A}$ and $\mathbf{C}$. K^* is the parameter that corresponds to the number chemical analytes present in the subset $\mathbf{C}$. Fig. 3 presents a block diagram of the KmP algorithm. In the figure, an unlabeled data set $\mathbf{X}$ is processed sequentially. First, K^* and m_D are estimated, and then they serve as the input parameters for the P -learning phase, which in turn computes a set of the class posteriors, $\mathbf{P}_\mathbf{X} = [P(l|\mathbf{r}_1^D), P(l|\mathbf{r}_2^D), P(l|\mathbf{r}_3^D), \dots, P(l|\mathbf{r}_N^D)]$ as the probabilistic representation of the class labels for each measurement. In the following sections, we introduce each of the three phases in detail.

Figure 3: The schematic diagram of the KmP algorithm, which consists of three sequential phases, K -learning, m -learning and P -learning phase respectively, and estimates the class posteriors of the data.

K –Learning Phase

The KmP algorithm starts with the K –learning phase, which estimates the number of detectable gases in the data. This first step is critical in comparison to previous approaches, since it allows the algorithm to process data without the need for prior information about the number of gases that might be present in the environment.

As shown in Fig. 3, the input data set $\mathbf{X}$ is divided in two subsets, namely $\mathbf{C}_{m_0}$ and $\mathbf{A}_{m_0}$, using an a-priori classification threshold m_0 . The a-priori classification threshold m_0 can be empirically set to be, e.g., about 30% of the maximum concentration level measured in $\mathbf{X}$. The final value of m is refined in the m –learning phase, once K^* is determined (Section 3.2). The goal of the data preprocessing block in the K –learning phase is to exclude all diluted measurements from the clustering process, and to consider only the measurements of analytes. As shown in the block diagram (Fig. 3), clustering analysis is performed over $\mathbf{C}_{m_0}$ with various K values. The corresponding results are evaluated using the silhouette criterion to determine an optimal K^* value [18].

Cluster analysis is performed using the Rodriguez-Laio algorithm [19] (hereafter called R-L) in the K –learning phase as well as subsequent phases. Thus, we first introduce the applied R-L clustering algorithm and the corresponding chosen pairwise distance (i.e., similarity) measure in Section 3.1.1 and Section 3.1.2. In Section 3.1.3, we give the details of the silhouette criterion evaluation, which is utilized for cluster validation, i.e., the determination of K^* .

The Rodriguez-Laio Clustering Algorithm

For the task of gas discrimination, the shortcomings of widely applied clustering techniques have to be taken into consideration. First of all, cluster with arbitrary shapes makes K-means not suitable for the task, since K-means can only detect hyper-spherical clusters [20]. Although DBSCAN (density-based spatial clustering of applications with noise) can be an alternative to handle arbitrary-shaped cluster [21], it requires users to manually set a density threshold to discard the data points with densities lower than this threshold as noise [19]. However, to set an appropriate density threshold often needs expertise knowledge. Other commonly used algorithms have some drawbacks that limit their applications in gas discrimination, such as the lack of robustness to noise (hierarchical clustering) [19,22], and the risk of relying on incorrect model assumption (Gaussian mixture methods) [19,20,22]. In contrast, a state-of-art clustering approach, namely the Rodriguez-Laio (R-L) algorithm, does not have the aforementioned disadvantages. In this work, it is used to carry out clustering analysis over an input subset $\mathbf{C}$. The R-L algorithm relies on the assumption that cluster centers are surrounded by neighbors with lower local density, and the centers are relatively distant from other data points with high local density [19]. Correspondingly, there are two key criteria in the algorithm: local density ρ and minimum-distance to higher density δ . First, the local density of a data point i is estimated as

$$\rho_i = \sum_{j \in N, j \neq i} e^{-(\frac{d_{ij}}{d_c})^2} \quad (1)$$

where d_c is the Gaussian kernel bandwidth and d_{ij} is the pairwise distance between data point i and j (see details in Section 3.1.2). Here, we modify the above kernel density estimator, by adopting a k -nearest neighbor approach [23], as follows:

$$\rho_i = e^{(-\frac{\sum_{j=1}^k d_{ij}}{k})} \quad (2)$$

where j belong to observation i 's k -nearest neighborhood, and d_{ij} is the pairwise similarity measure between observations i and j . A detailed description of distance measure d_{ij} will be given in Section 3.1.2. An appropriate k is chosen by $k = \lceil n \cdot N \rceil$, with $n < 1$, as the only free parameter of the clustering algorithm.

The second criterion, δ_i , is measured by computing the distance between the point i and its closest point with higher ρ , as follows:

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}) \quad (3)$$

For the data point with highest density, δ is set as $\delta_i = \max(d_{ij})$.

The concepts of ρ and δ are illustrated in Fig. 4, where two dense clusters are shown, and the cluster that point A belongs to has higher average density than the other cluster. Point A and point B have the highest ρ in each cluster under k -NN local density estimation given $k = 8$. The calculation of their ρ considers the pairwise distances between the point (A or B) and its 8 closest neighbors in the ellipse, which gives $\rho_A > \rho_B$. As an explanatory example, here we assume all other points have smaller local density than point A or B. Hence, for point B, its δ is its distance to point A, as point A is the closest point with higher local density.

Figure 4: An example of the calculation of ρ and δ for an artificial data set with k -NN local density estimation approach with $k = 8$. In this example, the cluster containing point A is denser than the cluster containing point B on average.

Fig. 5 shows how the k -NN approach differs from the Gaussian kernel approach. One can observe that in Fig. 5a, the two largest δ points (highlighted by a blue dashed box) given by the Gaussian approach have a clear gap in ρ values, whereas in Fig. 5b, the counterparts given by the k -NN approach have much closer (almost equal) ρ values in comparison. This example shows that, as an adaptive estimation of the local density [22], the k -NN approach allows the ρ value to be calculated in a local scale. It is expected to be beneficial to the clustering process in the sense that the k -NN approach manages to evaluate the importance of a measurement (with respect to ρ value) mainly according to its position in the local data structure, regardless of the density levels of other clusters. When the density level of clusters differ greatly, the points corresponding to the peaks of δ_i estimated by k -NN in each cluster may still have close ρ , because their ρ values are adapted by the average density level of their own neighbors, unlike the Gaussian kernel estimator using an universal d_c .

Recall the basic assumption of the algorithm, the clusters centers are recognized as points with both significantly large ρ and δ . Therefore, potential cluster centers are determined according to the indicator γ_i defined as follows,

$$\gamma_i = \rho_i \cdot \delta_i \quad (4)$$

The data points which correspond to the K^* largest γ_i are selected as cluster centers, where K^* is determined beforehand with a cluster validation method. After determining which data points are the cluster centers,

Figure 5: Comparison of Gaussian kernel approach and k -NN approach for local density estimation. The two approaches are applied to calculate the ρ and δ on the same data set, which has two classes. The ρ - δ plots show the (ρ, δ) values of the first 30 data points with the highest δ . The data used here is the same with Fig. 2. Note the difference between the scales of ρ values, which shows that the selected data are estimated to have much closer local density using k -NN approach.

the next step is to assign the remaining data points the same cluster labels as their nearest neighbors of higher ρ .

The computational complexity of the KmP approach mainly comes from the clustering process described above. The computations of ρ and δ require to calculate $N(N-1)/2$ pairwise distances and then to sort them in increasing orders for k -NN local density estimation, which contribute $O(N^2 \cdot \log N)$ complexity. The computation of γ and the final class label allocation procedure have complexity $O(N^2)$ and $O(N)$ respectively. Thus, the total complexity of the clustering process is no more than $O(N^2 \cdot \log N)$. Although the clustering process are carried out several times in the whole KmP algorithm, the most demanding computations only need to be performed once since the resulting distance matrix and the rank of ρ are reused in all the three phases of the KmP algorithm.

The Pairwise Distance measurement

Since the calculation of ρ and δ depends on the pairwise distance between measurements, it is critical to choose a distance measure that better quantifies the closeness, or similarity between two data points in the feature space. The distance measure d_{ij} is expected to be large when the data point i and j are from different classes, while d_{ij} should be relatively small when the data points are from the same class [24]. Given the instantaneous responses of the selected sensors as the features, the measurements from the same class distribute roughly along an ellipsoid (see Fig. 2a). For this reason, we opted for the cosine distance as the pairwise distance measure, which is defined as

$$d_{ij} = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|} \quad (5)$$

where $\mathbf{v}$ is the vector of extracted features from a given measurement. In our particular case, we use instantaneous sensor response $\mathbf{r}^D$ as features, where D is the number of considered sensors in the e-nose. Thus, $\mathbf{v}$ is a D -dimensional vector.

To demonstrate the applicability of the cosine distance, we compute the silhouette values of a set of acquired measurements. The silhouette value s_i is a metric of how similar the measurement i is to other measurements in its own class, when compared with measurements from a different class. It is defined by:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (6)$$

where a_i is the average distance from observation i to all other observations in the same class, and b_i is the minimum average distance from i to all the observations from different classes. A high silhouette value indicates that the observation is well-matched to its own cluster, and poorly-matched to neighboring clusters. If too many observations have a low or negative silhouette value, a clustering process is likely to fail because of the poor class separability, since the distance measure is not sufficient to reflect the characteristics of the data.

Fig. 11 shows examples of silhouette plots illustrating the separability of cosine distance and Euclidean distance. The silhouette plots show the descending ordered silhouette values of the observations in each class. Basically, the more observations have positive silhouette values, the better of the class separability is. One can tell that, in our case, cosine distance is a better descriptor of the pairwise distance, compared to the widely used Euclidean distance, which causes a lot of observations with negative silhouette values in class 1. Although, measuring with cosine distance still exhibits some observations with negative silhouette values. This is hard to avoid in data collected in open environments due to the presence of low-concentration observations that overlap in the feature space, as shown in Fig. 2.

Figure 6: Silhouette plots of a data set collected by mobile robot in an indoor arena (highly diluted observations are not included. See details in Section 4). The y-axis corresponds to the class label. **(a)**: Silhouette values using cosine distance. The measurements with negative silhouette values indicated by the dashed rectangle usually lie in the overlapping area; **(b)**: Silhouette values using Euclidean distance.

Silhouette Criterion for K^* Estimation

In the original R-L clustering algorithm, the number of clusters K^* is found along with the determination of the cluster centers (see Eq. 4). However, this critical step is carried out manually by visually identifying the data points with both large ρ and δ (or γ) in a so-called decision graph [19]. Although such procedure is straightforward, it is prone to errors (see the results in Section 5 for examples) and requires human intervention. In order to automatize the selection of the number of clusters, we utilize an unsupervised evaluation method to determine the most likely value of K^* . There are different assessment criteria for this task, such as the Calinski-Harabasz criterion [25], the Davies-Bouldin criterion [26], the silhouette criterion [18], among many others. Since we have used the silhouette criterion to measure the class separability for each individual measurement, with the same principle, it is natural to evaluate the fitness of a clustering results with

silhouette values. In general, a good clustering solution results to well-separated clusters that have larger between-cluster distance and small within-cluster distance [18]. The average silhouette value is an index to assess the overall clustering solution, which is calculated as follows:

$$\bar{s}(K) = \sum_{i=1}^N s_i / N \quad (7)$$

given the clustering result corresponding to K , where the individual silhouette value s_i is computed according to Eq. 6. The most suitable number of clusters K^* maximizes $\bar{s}(K)$.

m -Learning Phase

Using the number of classes K^* , the m -learning phase finds the classification threshold m_D that discriminates informative measurements of chemical analytes from those under clean air condition.

In the KmP algorithm, the classification threshold m_D is a means to handle highly diluted, low concentration measurements, which are found to be hard to discriminate [2,27]. In our case, they are also problematic for data clustering with the R-L algorithm because these diluted measurements form up a dense overlapping area in the feature space. To be precise, these diluted measurements lie on the corner region of the feature space, as shown in Fig. 2. In the KmP approach, these diluted measurements are assumed to correspond to the baseline responses in the absence of chemicals, and thus should be identified as an implicit class, namely clean air. However, as one can observe from Fig. 2, the boundary between the baseline responses and non-air chemical analytes in the feature space is rather ambiguous, so manually labeling these baseline responses as clean air is a matter of subjective judgment that relies on a-priori knowledge on the data collection process. For this reason, the m -learning phase learns this threshold from the acquired measurements directly. Besides the concentration level, it is neither possible, nor desirable to define another more suitable similarity measure that ensures efficient class separability between this dense, low concentration area and other measurement clusters of chemical analytes. Classification threshold is thus based on the concentration level of the data.

The implementation of the m -learning phase is illustrated in Fig. 7. The first step is to define an ascending sequence of classification threshold candidates, $\mathbf{m} = [m_1, m_2, m_3, \dots, m_T]$, to sample a series of subsets from $\mathbf{X}$, i.e., $\{\mathbf{C}_{m_1}, \mathbf{C}_{m_2}, \dots, \mathbf{C}_{m_T}\}$. Next, we run R-L clustering on each $\mathbf{C}_m$ to get the resulting class labels $\mathbf{L}_{K^*, m}$ for the following label disagreement analysis, which quantifies the discrepancy between the different clustering results. In the disagreement analysis, for each pair of subsets corresponding to (m_{t-1}, m_t) , the disagreement of the clustering results is quantified with the label disagreement indicator CI , defined as follows:

$$CI(s_1, s_2) = \sum_{i \in \mathbf{L}_{K^*, s_1} \cap \mathbf{L}_{K^*, s_2}} \omega_i \cdot \chi(\mathbf{L}_{K^*, s_1}(i), \mathbf{L}_{K^*, s_2}(i)) / N' \quad (8)$$

where ω is a weight proportional to the estimated concentration level I_c , and $\chi(x, y) = 1$ if $x = y$ and $\chi(x, y) = 0$ otherwise. $\mathbf{L}_s$ is the resulting class label vector under the classification threshold candidate s as $\mathbf{L}_{K^*, s} = [l_1, l_2, l_3, \dots, l_{N'}]$, where N' is the size of the union set of $\mathbf{C}_{s_1}$ and $\mathbf{C}_{s_2}$. The observations with high I_c are intuitively more informative because such observations lie on regions of the feature space that exhibit a higher degree of separability (e.g., Fig 2). The disagreement of labeling on such observations implies significant instability of the clustering results.

Consequently, we get a series of CI for each (m_{t-1}, m_t) for $t = 2, 3, \dots, T$, through which we can monitor the stability of the clustering results. As illustrated in Fig. 7, from left to right, as m increases, the disagreement of clustering results (i.e., $CI(m_{t-1}, m_t)$) tends to be decrease, whereas for the first few m_t the clustering results differ significantly from the next one. This is because the extracted subset $\mathbf{C}_m$ includes less diluted measurements given a higher m . To find the classification threshold, we are interested in the inflection point, if it exists, after which $CI(m_{t-1}, m_t)$ stabilizes to a low level. A straightforward strategy to

Figure 7: Illustration of the process in the m -learning phase. The colored rectangles are schematic representations of the clustering results $\mathbf{L}_{m_t}$ on the data $\mathbf{C}_{m_1}, \mathbf{C}_{m_2}, \dots, \mathbf{C}_{m_t}, \dots$. The white area in the bottom of the rectangle of $\mathbf{C}$ is the illustration of $\mathbf{A}$ discriminated by the classification threshold, and the blue and red color reflect the two class labels given by the clustering process.

determine m_D is to look for the m corresponding to the minimum derivative of $CI(m_{t-1}, m_t)$. However, it is possible that there could be another major decrement after the minimum derivative. In such case, the first decrement of CI is misleading since the cluster assignment has not been stabilized at that point. In order to handle this particular scenario while capturing the m that corresponds to the most significant decrement in $CI(m_{t-1}, m_t)$, the selection criterion for m_D is defined as follows:

$$CI(m_{t-1}, m_t) < \overline{CI}, \quad \forall m_t > m_D \quad (9)$$

which means m_D satisfies the requirement that all $CI(m_{t-1}, m_t)$ with $m_t > m_D$ are smaller than the average CI values.

P -Learning Phase

The P -learning phase assigns class posteriors to the measurements in $\mathbf{X}$. We apply a two-step strategy to estimate class posteriors, as follows:

1. Using the labels resulted from the clustering process, probabilistic classifiers are trained to compute pairwise probabilities between the discriminated substances $L_1, L_2, \dots, L_K$. A simple model is fit to

calculate pairwise probabilities between chemical substances and L_0 , where L_0 corresponding to clean air.

2. The pairwise probabilities of each measurement are coupled together using the method introduced in [13] and [28], to estimate the posterior of each class.

First, we describe how the pairwise probabilities between the target chemical compounds are determined. According to the operation principle of the clustering algorithm, a given observation is labeled according to the label of the closest measurement with higher local density. Such mechanism, however, does not directly reflect the degree of certainty on the class identity. We expect the indicator for posterior can quantifies how representative a measurement is. The δ , however, is not informative for this purpose. Similarly, ρ is the index for local density, which does not contain other information that is important to class separability, e.g., concentration level. For this reason, the class posteriors of the chemical analytes are not estimated from the clustering process. Instead, the pairwise probabilities among the samples are assessed by fitting a probabilistic model based on the estimated class labels. We thus utilize a probabilistic classifier trained with the outputs of the clustering process, with an underlying assumption that the clustering solution provides sufficient label information to the classifier. In this paper, a binary SVM classifier is used to compute the pairwise probabilities between chemical compounds since SVM can efficiently perform a non-linear classification. The SVM is trained by the labeled data obtained from the clustering on the data of $\mathbf{C}_{m_D}$, where m_D and K^* are provided from the last two phases.

Another factor contributing to the estimation of the class posteriors is the concentration levels of measurements. In a given gas discrimination problem, the certainty on the gas identity should also be related to the location of the measurement in the feature space (Fig. 2a). Those measurements located in areas that have a higher degree of separability should be labeled with higher confidence and vice versa. As the concentration levels increase, we observe that the overlapping of data points from different classes in the feature spaces decreases (Fig. 2a). As a result, the confidence on the identity of a chemical analyte should also increase. In other words, given a measurement $\mathbf{r}^D$, the class posterior of belonging to a chemical compound, $P(l|\mathbf{r}^D)$, tends to be high as long as the measurement is away from the overlapping areas in the feature space. To reflect this principle, we incorporate the concentration information to assess the uncertainty of assigning an observation to a particular class. Given a measurement of gas concentration I_c , its pairwise probabilities between a chemical analyte and clean air (i.e., L_0) is calculated with a simple model proposed in [13], as follows:

$$P_{l \vee L_0}(I_c) = 1 - e^{-\alpha_l I_c / \mu_l} \quad (10)$$

where the functional parameters, α_l and μ_l , denote the scale of the concentration levels in class l . The value of μ_l is the mean concentration level of class l . α_l is a class-dependent parameter to be fit, which should satisfy the relations $(P_{l \vee L_0}|I_c = \max(I_c)) \approx 1$ and $(P_{l \vee L_0}|I_c = \max(I_c)) \gg (P_{l \vee L_0}|I_c = m_D)$, where $\max(I_c)$ is the highest concentration level in class l . One simplistic way of selecting α_l is to set the class posteriors to a low value when I_c is close to the classification threshold, e.g., $(P_{l \vee L_0}|I_c = m_D) = 0.25$. Fig. 8 shows how Eq. 10 works. Given the ground-truth that all the observations illustrated belong to class 1, the values of the pairwise probability $P_{1 \vee 2}$ are rather fluctuant at low concentrations, indicating concentration levels play an important role in evaluating the fitness of the class prediction $P(l=1|\mathbf{r}^D)$.

Second, given in total $K + 1$ classes (K target chemical compounds and the implicit class air), the class posteriors $\mathbf{p} = [P(l = L_0|\mathbf{r}), P(l = L_1|\mathbf{r}), P(l = L_2|\mathbf{r}), \dots, P(l = L_K|\mathbf{r})]$ are computed by coupling the pairwise probabilities estimated by the binary classifiers. The computation of $\mathbf{p}$ is derived as solving a minimization problem. To begin with, $\mathbf{p}$ is supposed to minimize the cost function as follows [28]:

$$F = \sum_{i=0}^K \sum_{j=0}^K (P_{L_j \vee L_i} P_{L_i} - P_{L_i \vee L_j} P_{L_j})^2 \quad (11)$$

This cost function holds relying on the relation $(\sum_{i:i \neq j} P_{L_j \vee L_i} P_{L_i} - \sum_{i:i \neq j} P_{L_i \vee L_j} P_{L_j}) \approx 0$. According to

Figure 8: Pairwise class probability plot of the observations from class 1. The blue markers are $P_{1\vee 2}$ and the green and red dots denote $P_{2\vee L_0}$ and $1 - P_{2\vee L_0}$ respectively.

Wu and co-authors in [28], the problem in Eq. 11 is equivalent to following optimization formulation:

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2} \mathbf{p}^T \mathbf{Q} \mathbf{p} \\ \text{subject to} \quad & \mathbf{A} \mathbf{p} = 1 \end{aligned} \quad (12)$$

where $\mathbf{A} = [1, 1, 1, \dots, 1]$ is a $1 \times D$ vector and $\mathbf{Q}$ is the transition matrix that satisfies $\mathbf{Q} \mathbf{p} = \mathbf{p}$, given by

$$\mathbf{Q}_{ij} = \begin{cases} \sum_{s:s \neq i} P_{L_s \vee L_i}^2 + \alpha & \text{if } i = j \\ -P_{L_i \vee L_j} P_{L_j \vee L_i} + \alpha & \text{if } i \neq j \end{cases} \quad (13)$$

with $\alpha > 0$ to ensure $\mathbf{Q}$ is a positive definite matrix. The optimization problem in Eq. 12 can be solved using Lagrange multiplier method, which results in the following expression of $\mathbf{p}$:

$$\mathbf{p} = \mathbf{G} \mathbf{A}^T + [(\mathbf{A} \mathbf{Q}^{-1} \mathbf{A}^T)^{-1} \cdot \mathbf{A} \mathbf{Q}^{-1}]^T \quad \text{where} \quad \mathbf{G} = \mathbf{Q}^{-1} + \mathbf{Q}^{-1} \mathbf{A}^T (\mathbf{A} \mathbf{Q}^{-1} \mathbf{A}^T) \mathbf{A} \mathbf{Q}^{-1} \quad (14)$$

To sum up, after the clustering process, for each pair of classes in $\mathbf{L}$, a SVM probabilistic classifier is trained with labeled data $\mathbf{C}_{m_D}$. The outputs of these classifiers, the pairwise probabilities between chemical analytes, along with the pairwise probabilities given by Eq. 10, form up the transition matrix $\mathbf{Q}$, which will be used to compute $\mathbf{p}$ according to Eq. 14.

Data Set Description

We carried out gas discrimination on data sampled with commercial MOX sensor arrays in different open environments. The data are referred to as robot arena (RA), outdoor courtyard (OC), and a wind tunnel (WT3 and WT4). A summary of the main characteristics of each data set is given in Table 1, and a brief description covering the experimental configurations and scenarios is given below.

Table 1: Summary of the experimental configurations of the three gas data sets used in this work.

Data set	Sensors	Gases	Sampling frequency
Robot Arena (RA) [28]	MICS-2610, MICS-2710, MICS-5121, MICS-5135, MICS-5521×2	ethanol, propanol	4 Hz
Outdoor Courtyard (OC) [13]	TGS2600, TGS2602, TGS2611, TGS2620	ethanol, acetone	4 Hz
Wind Tunnel (WT3, WT4) [10]	TGS2611, TGS2612, TGS2610, TGS2600, TGS2602×2, TGS2620×2	acetylene, toluene, benzene, ethylene methanol	100 Hz (original) 50 Hz (sub-sampled)

The robot arena (RA) data set was collected with a mobile robot exploring in a 5 m × 5 m unventilated room with an array of 5 MOX sensors (Fig. 9a). The data set consists of 6 independent single-source trials, each of which has a 1800 seconds duration. In each trial, the target chemical compound measured is either propanol or ethanol, which is released into the air by pumps placed on the ground. The robot followed a predefined spiral path at a speed 0.05 m/s, and was programmed to stop at specified measurement spots to collect data for 30 seconds [29].

The outdoor courtyard (OC) data set was collected by a robotic platform, equipped with 4 MOX sensors in an outdoor 9 m × 7 m garden surrounded by nearby buildings (Fig. 9b). The data set consists of 4 single-source trials, which have a duration of 3400 seconds. In each trial, the target analyte acetone or ethanol is evaporated from open plastic containers. The robot was remotely controlled to move in a random trajectory in the courtyard with a speed of around 0.14 m/s, and stopped at different way-points to collect measurements for about 30 to 60 seconds per way-point [13].

In addition, in the same environments of the data sets RA and OC, several dual-source experimental runs were conducted, in which two pumps released different gases at the same time. In this paper, the corresponding acquired trials are collectively called gas mixture data sets.

The wind tunnel data set was collected by 9 sensor array modules equipped with 8 MOX sensors positioned inside a wind tunnel (Fig. 9c) [10]. In the original data set presented in [10], the sensor arrays were placed to take measurements at a variety of distances to the gas source, and there are up to 10 gaseous substances, released in independent experiments that aim to resemble the complicated conditions of real environments. In each experiment, the e-noses were placed at predefined positions, and exposed to clean air at first. Then, one of the 10 chemical compounds was released into the tunnel for 20 seconds. During this period, the chemical analyte circulated throughout the wind tunnel under a preset airflow strength for 3 minutes. At last, the gas source was removed and the environment was ventilated. We extract two subsets from the extensive original data to test the *KmP* approach with more challenging multi-compound gas discrimination tasks. The criteria of the extraction considers two aspects, the sampling positions and the number of substances. First, the measurements sampled at the following positions are considered: the fifth row of the columns P1, P2, P3, P4, P5 and P6 (see Fig. 9c). The considered sensory array was always at the middle row in the chamber, which is most likely to be fully exposed to the flow of airborne gases during the experiment. Second, we extract a 3-class and a 4-class subsets, which are named as WT3 and WT4 respectively to indicate the number of chemical compounds in the corresponding data sets. WT3 contains acetylene, methanol and benzene. WT4 includes acetylene, toluene, benzene, and ethylene. We examine the class separability of different combinations of the substances using silhouette values (Eq. 6 and Eq. 7). We find that some combinations yields very poor class separability (with average silhouette values $\bar{s} < 0.2$), which are not suitable for testing cluster-based discrimination methods because in such cases the sensors show cross-selectivity to some of

the substances. There are some combinations have relatively good class separability ($\bar{s} > 0.45$), which are not chosen either since they are not challenging enough as test sets. WT3 and WT4 have $\bar{s} = 0.4246$ and 0.3227 respectively, which are feasible for cluster-based methods, but the class separabilities are not perfect due to the environmental conditions. The changing environmental condition in WT3 and WT4 is the wind flow, which is generated by the outlet fan as shown in Fig. 9c. 3 levels of artificial wind flow strengths, namely “Small”, “Medium”, “Strong”, were created by setting the ventilation fan to rotate at three different constant rotational speeds. The wind condition “Small” corresponds to a wind speed of 0.1m/s, “Medium” corresponds to 0.21 m/s, and “Strong” corresponds to 0.34 m/s. In this paper, unless otherwise specified, WT3 contains the data of 3 gaseous compounds under all wind flow conditions, and WT4 includes 4 classes under the “Strong” wind flow. The sensors used in wind tunnel data set have 100 Hz sampling frequency. To reduce the computational complexity, both WT3 and WT4 are sub-sampled from the original data set with frequency 50 Hz.

Figure 9: The data collection environments of the data sets RA, OC, WT3 and WT4. (a): the robot in a large room (the data set RA) [29]; (b): the robot in a courtyard (the data set OC) [13]; (c): the wind tunnel test-bed facility (the data sets WT3 and WT4). Image is adapted from [10].

Results

K –Learning Evaluation

In the KmP algorithm, the cluster centers are automatically determined by applying the silhouette criterion. This strategy is compared against the decision graph suggested in the original R-L algorithm [19], the Calinski-Harabasz criterion and the Davies-Bouldin criterion. The Calinski-Harabasz criterion is based on a ratio of overall average between-cluster distance variance and within-cluster variance [25]. The Davies-Bouldin criterion evaluates the cluster validity based on a defined distance between two clusters, and finds the optimal clustering solution that corresponds to the smallest average pairwise between-cluster distance [26].

As shown in Fig 10, the number of clusters K^* is found correctly in all data sets. K^* s correspond to the maximum average silhouette value (see the results of using the silhouette criterion in Fig. 10d, Fig. 10h, Fig. 10l and Fig. 10p), whereas the decision graphs γ –plots do not provide very clear indication in the 4-class case (WT4, see Fig. 10m). In comparison, the Calinski-Harabasz and Davies-Bouldin criterion are less reliable in the cases of multi-class data sets (WT3 and WT4, see Fig. 10j, Fig. 10k, Fig. 10n and Fig. 10o).

Figure 10: Cluster validity estimation on the data sets. (a) to (d): the data set RA has 2 classes of chemical substances, indicated successfully by all criteria; (e) to (h): the data set OC has two classes, indicated successfully by all criteria; (i) to (l): the data set WT3 has 3 classes, indicated successfully only by the silhouette criterion; (m) to (p): the data set WT4 has 4 classes, indicated successfully only by the silhouette criterion.

m–Learning Evaluation

We validate our strategy in the *m*–learning phase by looking into the class separability of the measurements that correspond to the **C** subset. As explained in Section 3.2, the data clustering should be performed on measurements of chemical compounds, so that the diluted samples measured in clean air will not impair the clustering process. For this reason, the classification threshold m_D is learned to extract the subset $\mathbf{C}_{m_D}$ for a better class separability than the universal set **X**.

One expected outcome of extracting data with a classification threshold is the improvement on class separability. To demonstrate this, we take a series of subsets $\mathbf{C}_m$ from $\mathbf{X}$ by applying various classification thresholds $\mathbf{m} = [m_1, m_2, m_3, \dots]$, and evaluate the silhouette values of the measurements in each subset. An example case is shown in Fig. 11a. As the value of m increases, the class separability of the extracted subset $\mathbf{C}_m$ (indicated by its average silhouette value) is consistently enhanced. This phenomenon implies that the clustering solution significantly improves after the classification threshold has been set.

Figure 11: Average silhouette value and CI under different m . **(b)**: The average silhouette value of the resulting cluster solution under various classification threshold candidates m for the data set RA. The error bars represent the variance of the overall silhouette values of the measurements in each subset. The dashed box points to the maximum increment of the average silhouette value. **(b)**: m_D is found in the m -learning phase is in the range indicated by the dashed box in Fig. 11a.

Fig. 12 shows how the classification thresholds are found in the m -learning phase for the four data sets.

In each sub-figure, one can observe that label disagreement CI (Eq. 8) peaks in the relatively early stages, which implies that the increment of m leads to substantially different clustering solutions on the same data. What follows on is that CI decreases and stays close to zero, which means the clustering assignment does not drastically differ. Recall our criterion described in Section 3.2, where we defined that the classification threshold m_D corresponds to the last turning point of CI . As expected, in each case shown here, m_D is always captured at the m where CI has an obvious decline and exhibits no major change afterwards. One detail in the implementation of the m -learning phase is the choice of the range and step size of the sequence of classification threshold candidates $\mathbf{m}$, which is a trade-off between computation cost and precision that is up to the user to decide case by case. For example, in Fig. 12a, the number and range of the classification threshold candidates are obviously different from other data sets. We suggest at least set 5 pairs of (m_{t-1}, m_t) in practice.

In this work, the classification threshold is based on the estimated gas concentration I_c computed with readings from the sensor array using the method proposed in [13]. In data set RA and OC, the gas concentrations were also recorded with a PID, which is calibrated and therefore more precise. Based on this relatively accurate concentration information, we try to approximate the concentration level that corresponds to the classification threshold. Considering the PID responses of all the measurements in $\mathbf{A}_{m_D}$, we estimate that, approximately in RA data set the found m_D corresponds to 75.7350 ppm, and in OC data set, the m_D corresponds to 7.0176 ppm.

Figure 12: The classification threshold found by monitoring the CI index. The arrow points to the classification threshold. (a): the data set RA; (b): the data set OC; (c): the data set WT3; d): the data set WT4.

Classification Performance

This section is devoted to assess the performance of the proposed KmP approach in terms of the following aspects: (1) the robustness with respect to the selection of its only free parameter neighborhood size

(Section 5.3.2); (2) its performance compared to commonly used clustering algorithms (Section 5.3.3); (3) its robustness with respect to changing environmental conditions (Section 5.3.4). The metrics used for the above evaluation are described in Section 5.3.1.

Experimental Validation

We apply the so-called classification-oriented cluster validity to evaluate the performance of the data clustering procedure in P -phase. The classification-oriented evaluation quantifies the degree to which the predicted class labels deviate from the actual ones in the ground-truth. In this paper, we use classification rate and Cohen’s kappa in the comparison, which are defined as follows:

- Classification rate CR is also known as exact match ratio, or overall success rate. It is defined as

$$CR = \frac{\text{\#correct predictions}}{\text{\#total predictions}} \quad (15)$$

- Cohen’s kappa CK is used to measure the agreement of the categorical items (i.e., the assigned labels) between ground-truth and the prediction by the classifier. Generally, it is thought to be a more robust measure than the extract match ratio, because CK is offset by considering the agreement occurring by chance. The equation for CK is as follows:

$$CK = \frac{p_o - p_e}{1 - p_e} \quad (16)$$

where p_o is the overall agreement rate among the prediction and the ground-truth, and p_e is the hypothetical probability of chance agreement, which is calculated as the classification rate obtained by randomly assigning the labels to the observations. For a good classification performance, we expect CK to be close to 1.

Robustness to Parameter Selection

To demonstrate the robustness of the algorithm with respect to the only functional parameter neighborhood size k , given by $k = \lceil n \cdot N \rceil$, we conduct gas discrimination using different values of n . For each value of n , K -fold cross validation is used to show the variation in the clustering performance. Since the data sets RA, OC, WT3 and WT4 cover a board variability with respect of the number of classes, the type of gases and the sensing conditions, the detailed set-ups of folds for each data set differ in a few subtle aspects, which are introduced as follows:

- For the data set RA and OC, the folds are independent experimental trials.
- For the data set WT3 and WT4, the folds are sampled subsets corresponding to three different airflow conditions (“Small”, “Medium” and “Strong”).

The results are illustrated in Fig. 13. For the data set RA, we find that except for $n = 7\%$, the algorithm maintains high performance (on average around 95% classification rate) for different values of n . For the data sets OC, WT3 and WT4, the algorithm is also robust in a wide range, and favors $5\% < n < 20\%$. Considering all cases, it is safe to conclude that the algorithm shows a rather stable performance when n is set within the range of 10% to 20%. A relatively small value of n is also good for computational efficiency, in the summation operation Eq. 2.

Figure 13: The robustness evaluation of the algorithm with respect to the parameter n using classification rate CR and Cohen's kappa CK . (a): the data set RA; (b): the data set OC; (c): the data set WT3; (d): the data set WT4.

Performance Comparison

In order to validate the classification capability of the KmP algorithm, we compare our approach with other unsupervised algorithms, including the original R-L clustering algorithm (using Gaussian kernel instead of k -NN approach for density peak detection), K-means [30], a Gaussian mixture model [31] and agglomerate hierarchical clustering (using complete linkage) [20]. For a given data set $\mathbf{X}$, we are interested in the classification accuracy on the subset $\mathbf{C}$ that corresponds to the chemical analytes. For this reason, we first run the m -learning phase to extract the subset $\mathbf{C}_{mD}$, which will be the test set to be processed by the all the clustering algorithms. The clustering results are analyzed with respect to classification rate CR and Cohen's kappa CK . As K-means and Gaussian mixture model are heuristic algorithms, their performances depend on initial conditions. For this reason, there is not guarantee that a single run of K-means (or Gaussian mixture model) will provide its optimal result. In practice, K-means and Gaussian mixture model are repeated multiple times with different initial centroids, and one of the clustering solutions is selected with an unsupervised criterion as the final result. For example, for K-means, the criterion is to choose the result that yields the least sum of point-to-centroid distances. In our experiment, K-means and Gaussian

mixture model are repeated for 10 times and 100 times respectively. Other algorithms to be compared are deterministic, so they are run only once. For the R-L algorithm, its free parameter is set according to the rule of thumb suggested in [19]. The comparative results are summarized in Table 2 and Fig. 14.

Table 2: The summary of classification results using our approach and classic clustering algorithms.

	RA		OC		WT3		WT4	
	CR	CK	CR	CK	CR	CK	CR	CK
The KmP algorithm	0.9853	0.9703	0.9639	0.9212	0.9881	0.9822	0.9400	0.9200
The R-L algorithm	0.9802	0.9599	0.9197	0.8198	0.8779	0.8187	0.9321	0.9095
K-means (KM)	0.9797	0.9591	0.9125	0.8042	0.9889	0.9834	0.9178	0.8904
Gaussian mixture model (GMM)	0.8216	0.6493	0.4963	-0.0932	0.9672	0.9508	0.9305	0.9072
Hierarchical clustering (HC)	0.8103	0.6013	0.6764	0.1529	0.8007	0.7071	0.3767	0.1651

Figure 14: The box plots of classification performances of each algorithm for the four data sets. The specific values of CR can be found in Table 2. The ends of the whiskers represent the 9th percentile and the 91st percentile, and the ends of the boxes represent 25th percentile and 75th percentile. The red cross and line represent outliers and median respectively. The KmP , the R-L and hierarchical clustering algorithm always appear with a single line because their results only contain one solution (no replication performed). (a): the data set RA; (b): the data set OC; (c): the data set WT3; (d): the data set WT4.

For all data sets, the KmP approach has a constantly good performance with relatively high classification rate (CR) and per-class accuracy (estimated by CK). K-means is the best performing among other algorithms, which has close accuracy with the KmP approach in all data sets. The results also show that hierarchical clustering and Gaussian mixture model seem to be disadvantageous for the data sets RA and

OC. In particular, for the data set OC, the Gaussian mixture model results in $CK = -0.0932$, which means the classification solution totally fails in each class. As is known, Gaussian mixture model clustering depends on the assumption that data fits a certain multivariate mixture of Gaussians. However, from the feature spaces of the RA and OC in Fig. 15a and Fig. 15b respectively, one can observe that the data sets do not fit Gaussian models very well, which might explain why it is not a suitable tool in these cases. Hierarchical clustering fails in all cases most likely because it is very sensitive to the outliers and noise. We also found that, if only considering measurements of high concentration levels (higher than the found classification threshold), hierarchical clustering can reach 100% CR with the data sets RA and OC. Such different behaviors implies a lack of robustness to the overlapping measurements. Another remark of our clustering algorithm is the improvement made by the adaptive density peaks detection. In comparison, the R-L algorithm, using the Gaussian kernel approach (Eq. 1), does not perform as good as the KmP algorithm in the data sets OC and WT3. The variations due to the repetitions for K-means and Gaussian mixture model are reported in Fig. 14. We observe that in the case of WT3 (Fig. 14c) and WT4 (Fig. 14d), the majority of repetitions of Gaussian mixture model do not give the optimal results. K-means exhibits no variation in the repetitions on RA, OC and WT3, but half of its repetitions on WT4 do not converge to a local optimum (which is not the final selected result). Such variations indicate that the repetitions are necessary when K-means and Gaussian mixture model are used. However, performing repetition and selection will increase the computational cost, which is a disadvantage of K-means and Gaussian mixture model compared to other deterministic algorithms.

Fig. 15a shows the clustering result illustrated in the feature space, where we can observe that misclassifications are mostly those observations with small cosine distances to other classes, and the non-spherical clusters are successfully detected. For the case of RA and OC, misclassifications only occur at low-concentration regions. In the data set WT3 (Fig. 15c), some misclassifications also occur at an overlapping region with higher gas concentration. For the data set WT4 (Fig. 16a), we find misclassifications at a high-concentration, isolated region (i.e., the misclassifications in the ellipse). These measurements are sampled in close proximity to the gas source in the wind tunnel. We find that some of the measurements obtained at this position have the highest concentration levels, but in the feature space, their positions are isolated from other representative measurements belonging to the same analyte. Since the KmP algorithm uses cosine distance, the data points in the ellipse (in fact belonging to the analyte 2) have closer distance with the analyte 3 than the analyte 2, which causes the misclassification. In Fig. 16b, this can be seen clearly from another view.

(a)

(b)

Figure 15: The clustering results of the data sets RA, OC, and WT3. The misclassified observations are marked with black crosses. (a) The data set RA; (b) The data set OC; (c) The data set WT3.

Robustness to Changing Airflow

A major challenge for the *KmP* algorithm and many other gas discrimination approaches is that the changing environmental conditions affect the performance. The data sets WT3 and WT4 were taken under three wind speeds in a wind tunnel. Such experimental set-ups give us the chance of testing the *KmP* approach under various combinations of airflow conditions.

We simulate the “changing” airflow by combining at least two data sets of different wind speed. In Table 2 we provide the results for the data of various combinations of all three wind flow conditions (“Small”, “Medium” and “Strong”). The results suggests that our approach can handle the data in the environment with dynamic air flow. To be specific, the *KmP* approach performs well on the combinations “Small” & “Medium” and “Medium” & “Strong”. In the combination “Small” and “Strong”, we observe a degradation of the classification accuracy, which is probably due to the sudden transition in term of the wind flow.

Figure 16: Two views of the feature space of WT4. The clustering achieves high accuracy, but some misclassifications occur with high concentration levels. The ellipses in both sub-figures encircle the same region of misclassifications. The ellipses in both sub-figures encircle the same region of misclassifications.

Table 3: The summary of classification results using our approach under various airflow conditions.

Wind flow condition	WT3		WT4	
	<i>CR</i>	<i>CK</i>	<i>CR</i>	<i>CK</i>
Small	0.9852	0.9777	0.9764	0.9685
Medium	0.9775	0.9662	0.9689	0.9585
Strong	0.9563	0.9344	0.9400	0.9200
Small & Medium	0.9836	0.9754	0.9002	0.8667
Small & Strong	0.8374	0.7581	0.7476	0.6632
Medium & strong	0.9715	0.9572	0.9297	0.9062
Small & Medium & Strong	0.9881	0.9822	0.9380	0.9173

Class Posterior Estimation

In this section, we validate the posterior estimation procedure in P -learning phase. In the last section, on the basis that the clustering procedure achieves high success rate in the class label assignment, the labeled data are used to train a probabilistic SVM classifier to learn class posteriors. The parameter α_l in Eq. 10 is set to fit $(P_{l|V_{L_0}}|I_c = m_D) = 0.25$.

Fig. 17 and Fig. 18 show examples of the predicted posteriors. In both figures, the gas concentrations are shown as I_c in arbitrary unit. Since the actual gas identity is unknown to the algorithm, it is not possible to calculate the ppm value from the MOX sensor responses. Instead, we show the uncalibrated estimates, which are proportional to the real concentration (within the limits of the sensor characteristics). As expected, the measurements around the classification threshold (the black dashed line) are assigned with low posteriors. This implies that the estimation of $P(l|\mathbf{r}^D)$ takes the concentration level of $\mathbf{r}^D$ into account. The results reflect how the uncertainty decreases as the concentration levels rises away from the baseline. In the time series of sensor responses, we also find that the misclassifications, are isolated by a few measurements of small posteriors $P(l|\mathbf{r}^D) \leq 0.4$ (see the measurements in the gray boxes). Fig. 19 is a plot of class posterior time series of about 26 seconds in one trial of WT3. In this period, the sensors were exposed to acetylene. The P -learning phase assigned very high posteriors (class 1) to almost all corrected labeled measurements. One also can observe that the posteriors of class 1 fluctuate with the concentration levels. Again, it is because when the concentration recovered to the baseline level, the certainty of the measurements' identity decreases accordingly. The misclassified measurements are mostly assigned with relatively small posteriors while the measurements with correct labels are very likely to have high posterior values. Fig. 20 shows a 40 seconds posterior time series in one trial of WT4. During this period, a few misclassifications occur at high concentration levels, making the corresponding posteriors have values close to 1. This unwanted result comes from the measurements collected near the gas source (encircled by the ellipse in Fig. 20, Fig. 16a and Fig. 16b. See Section 5.3.3 for detailed explanation). The rest misclassifications are assigned with relatively small values as expected.

We also validated the P -learning phase with dual-source data sets. In contrast to the data sets RA and OC, we do not have the ground-truth of the chemical identity of each measurement in the dual-source data sets, i.e., class labels for each observation are lacking. For this reason, the evaluation is conducted in a qualitative manner. For example, Fig. 21 presents the spatial distribution of the observations acquired in a dual-source experiment with set-ups of the data set RA. In this case, we find that the spatial distribution of the class posteriors can provide considerable information: one can notice that our approach assigns low posteriors to the observations of low concentration levels, and makes class predictions correlated to the actual locations of the gas sources, i.e., the observations close to one gas source are predicted to have different label from the observations close to another gas source. At positions away from the gas sources, the measurements are recognized as the clean air with high confidence.

Figure 17: Visualization of class posterior for the data set RA. The colors green and red represent that the measurements are identified as propanol or ethanol, and yellow corresponds to clean air. The color shades denote the posterior values. **(a)**: The overall posterior time series; **(b)** and **(c)**: Zoom-in plots of the region A and B in **(a)**. Note that the posteriors are relatively small when the gas concentrations are close to the classification threshold.

Figure 18: Visualization of class posterior in the time series for the data set OC. The color green and red represent that the measurements are identified as acetone or ethanol, and yellow corresponds to clean air.

Figure 19: Visualization of a class posterior time series of about 26 seconds (1320 time steps) in WT3. During this period, the analyte that sensors were exposed to is acetylene. The plot at the top is the time series of the class posteriors for each class, in which red squares represent the posteriors assigned to the misclassifications. The plot at the bottom is the corresponding time series of concentration levels.

Figure 20: Visualization of a class posterior time series of 40 seconds (2000 time steps) in WT4. During this period, the sensors were exposed to toluene. The ellipse encircles the same measurements in Fig. 16a and Fig. 16b.

Conclusions and Discussion

In this paper, we presented the *KmP* algorithm, an unsupervised approach for gas discrimination. Compared to existing solutions for gas discrimination in open environments, the *KmP* algorithm can handle an unknown number of chemical compounds and does not need training with labeled data to detect the presence of different substances. Such capability is particularly useful for some real-world applications, in which the lack of information in target analytes and the presence of unknown interferences often make supervised learning not feasible. The *KmP* algorithm deduces the number of clusters from the data, and learns a classification threshold to discriminate chemical analytes from samples acquired in clean air conditions, which allows the clustering to perform on the data with favorable class separability. Learning this threshold allows to deal with unbalanced data sets (with respect to concentration levels), which are commonly observed when using open sampling systems. The data clustering in the *KmP* algorithm employs an adaptive local density estimation approach, which is an improvement of the original R-L clustering algorithm. Our clustering procedure has more stable performance on multi-class data sets in comparison. The output of the *KmP* algorithm are the class posteriors for both chemical substances and clean air, which, in particular, considers the correlation between the gas concentration and class separability.

For the purpose of methodology evaluation, we conducted in total 14 classification experiments to test the discrimination accuracy, 4 cross-validations to analyze the impact of the only free parameter on the clustering performance, and a qualitative clustering experiment on a dual-gas data set without ground-truth. In these experiments, the data sets include up to 4 different chemical substances and have a variety of experimental set-ups, ranging from a wind tunnel to an outdoor courtyard. There are three highlights in the results. First, from the comparative performance evaluation with the R-L algorithm, K-means, Gaussian mixture model and hierarchical clustering, we observed that our proposed approach has a stable and good gas discrimination performance. Second, the results of cross-validations suggest the *KmP* approach is robust to its free parameter. Third, the result of the qualitative experiment shows that class posterior estimation in *P*-learning phase can provide additional information that allows the user to have better understanding

Figure 21: Spatial distribution of concentrations of the observations acquired in an experiment with set-ups of the data set RA. The observations are colored according to their class posteriors, The z-axis corresponds to the concentration level of the observations, and the red and green circle markers indicate the positions of two gas sources respectively.

of the class labels obtained, which is also useful information for other subsequent tasks like gas distribution mapping and gas source declaration for the applications in mobile robotic olfaction [32]. In summary, we demonstrate that the *KmP* approach is feasible and effective in gas discrimination in open environments, which is in particular helpful for those applications that supervised or semi-supervised methods can be hardly applied due to the lack of training data. Moreover, the unsupervised approach refers to the effort that we try to understand the underlying data structure rather than fit a specific, pre-decided one.

Although our approach has the mentioned advantages on coping with data in open environments over some classical clustering algorithms, it has a few limitations. The performance of the cluster analysis, to a large extent, relies on the class separability of the extracted features. The features and the corresponding pairwise distance measure, play a crucial role in determining the number of classes and the clustering process. However, the feature extraction used in this work has no promise on good performance if the extracted features yield poor separability between different classes. This issue becomes evident with the classification on data set WT4. In that case, the cosine distance measure determined a close proximity relationship between non-alike measurements, as shown in Fig. 16b.

As future work, we would like to further develop the *KmP* approach into an on-line gas discrimination algorithm. The current *KmP* approach is designed to be used off-line, which means that the data set to be discriminated does not change when it is processed through the three phases. If the discrimination model is used to process a new class of substance that has not appear before, the measurements correspond to the new class would be mistakenly recognized as one of the previously detected classes. An on-line version *KmP* approach will overcome such limitation, and benefit some real-time robotic olfaction applications, such as environmental surveillance and inspection. In particular, the outputs of *KmP* approach can be coupled to the path planning strategy of the robots to optimize the collection of information and robot movement. Another potential improvement could come from exploiting the temporal information of the data. Currently, the discrimination process in the *KmP* algorithm only considers the spatial correlation between the data points in the feature space. but does not incorporate temporal information from the time-series of signals. Thus, in the future, we will investigate combined distance measures that consider the feature space domain as well as temporal domain. An related technique called ensemble clustering will also considered in our future research. For the same data set, this technique reconciles clustering solutions coming from several runs using different distances measures. Ensemble clustering has been found to be less sensitive to outliers and variations in data, and can outperform clustering process only relying on a single distance measure [33]. Moreover, to speed up the *KmP* approach, parallel computing could be applied. First, in the *K*-learning phase and *m*-learning phase, the clustering algorithm is executed several times, which is suitable to be performed simultaneously. Second, the main contribution of computational complexity, the computation of pairwise distances for all data points, can be split into several parts and solved in parallel. This potential improvement is rather important for on-line applications, in which an ongoing sampling process will keep introducing new data into $\mathbf{X}$, and therefore the calculation of pairwise distances has to be carried out frequently. In such scenario, distributing large and increasing data to multiply computation resources will reduce the time and memory consumption.

Acknowledgement

This work was funded in part by the H2020-ICT project SmokeBot (645101).

References

- [1] M. Trincavelli, M. Reggente, S. Coradeschi, A. Loutfi, H. Ishida, A. Lilienthal, Towards environmental monitoring with mobile robots, 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. (2008). doi:10.1109/iros.2008.4650755.

- [2] L. Dentoni, L. Capelli, S. Sironi, R. Rosso, S. Zanetti, M. Torre, Development of an electronic nose for environmental odour monitoring, *Sensors*. 12 (2012) 14363–14381. doi:10.3390/s121114363.
- [3] J.G. Monroy, J. Gonzalez-Jimenez, Gas classification in motion: an experimental analysis, *Sensors and Actuators B: Chemical*. 240 (2017) 1205–1215. doi:10.1016/j.snb.2016.09.013.
- [4] M. Trincavelli, S. Coradeschi, A. Loutfi, Odour classification system for continuous monitoring applications, *Sensors and Actuators B: Chemical*. 139 (2009) 265–273. doi:10.1016/j.snb.2009.03.018.
- [5] J.G. Monroy, J. Gonzalez-Jimenez, Odor classification in motion: how fast can the e-nose go?, in: 16th International Symposium on Olfaction and Electronic Nose (ISOEN), Dijon, 2015.
- [6] J.W. Gardner, P.N. Bartlett, *Electronic Noses: Principles and Applications*, Oxford University Press, Oxford, 1999.
- [7] A. Szczurek, B. Krawczyk, M. Maciejewska, VOCs classification based on the committee of classifiers coupled with single sensor signals, *Chemometrics and Intelligent Laboratory Systems*. 125 (2013) 1–10. doi:10.1016/j.chemolab.2013.03.007.
- [8] R. Gutierrez-Osuna, A. Gutierrez-Galvez, N. Powar, Transient response analysis for temperature-modulated chemoresistors, *Sensors and Actuators B: Chemical*. 93 (2003) 57–66. doi:10.1016/s0925-4005(03)00248-x.
- [9] C. Distanto, M. Leo, P. Siciliano, K.C. Persaud, On the study of feature extraction methods for an electronic nose, *Sensors and Actuators B: Chemical*. 87 (2002) 274–288. doi:10.1016/s0925-4005(02)00247-2.
- [10] A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov, R. Huerta, On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines, *Sensors and Actuators B: Chemical*. 185 (2013) 462–477. doi:10.1016/j.snb.2013.05.027.
- [11] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, A. Vergara, R. Huerta, Chemical discrimination in turbulent gas mixtures with MOX sensors validated by gas Chromatography-mass spectrometry, *Sensors*. 14 (2014) 19336–19353. doi:10.3390/s141019336.
- [13] V.H. Bennetts, E. Schaffernicht, V.P. Sese, A.J. Lilienthal, M. Trincavelli, A novel approach for gas discrimination in natural environments with open sampling systems, *IEEE SENSORS 2014 Proceedings*. (2014). doi:10.1109/icsens.2014.6985437.
- [12] M. Trincavelli, Gas discrimination for mobile robots, *KI - Künstliche Intelligenz*. 25 (2011) 351–354. doi:10.1007/s13218-011-0104-0.
- [14] L. Eusebio, L. Capelli, S. Sironi, Electronic nose testing procedure for the definition of minimum performance requirements for environmental odor monitoring, *Sensors*. 16 (2016) 1548. doi:10.3390/s16091548.
- [15] P. Jia, T. Huang, S. Duan, L. Ge, J. Yan, L. Wang, A novel semi-supervised electronic nose learning technique: M-Training, *Sensors*. 16 (2016) 370. doi:10.3390/s16030370.
- [16] F.-M. Schleif, B. Hammer, J.G. Monroy, J.G. Jimenez, J.-L. Blanco-Claraco, M. Biehl, et al., Odor recognition in robotics applications by discriminative time-series modeling, *Pattern Anal Appl Pattern Analysis and Applications*. 19 (2015) 207–220. doi:10.1007/s10044-014-0442-2.
- [17] S. Brahim-Belhaouari, M. Hassan, N. Walter, A. Bermak, Advanced statistical metrics for gas identification system with quantification feedback, *IEEE Sensors Journal*. 15 (2015) 1705–1715.
- [18] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*. 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.
- [19] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, *Science*. 344 (2014) 1492–1496. doi:10.1126/science.1242072.

- [20] R. Xu, D.C. Wunsch, Recent advances in cluster analysis, *International Journal of Intelligent Computing and Cybernetics*. 1 (2008) 484–508. doi:10.1108/17563780810919087.
- [21] G.-Z. Wang, G.-L. Wang, Improved fast DBSCAN algorithm, *Journal of Computer Applications*. 29 (2009) 2505–2508. doi:10.3724/sp.j.1087.2009.02505.
- [22] X.-F. Wang, Y. Xu, Fast clustering using adaptive density peak detection, *Statistical Methods in Medical Research*. (2015). doi:10.1177/0962280215609948.
- [23] J. Ding, S. Shah, A. Condon, densityCut: an efficient and versatile topological approach for automatic clustering of biological data, *Bioinformatics*. 32 (2016) 2567–2576. doi:10.1093/bioinformatics/btw227.
- [24] K. Brudzewski, S. Osowski, K. Wolinska, J. Ulaczyk, Smell similarity on the basis of gas sensor array measurements, *Sensors and Actuators B: Chemical*. 129 (2008) 643–651. doi:10.1016/j.snb.2007.09.050.
- [25] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics - Theory and Methods*. 3 (1974) 1–27. doi:10.1080/03610927408827101.
- [26] D.L. Davies, D.W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-1 (1979) 224–227. doi:10.1109/tpami.1979.4766909.
- [27] L. Capelli, L. Dentoni, S. Sironi, R.D. Rosso, The need for electronic noses for environmental odour exposure assessment, *Water Science & Technology*. 69 (2014) 135. doi:10.2166/wst.2013.544.
- [28] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, *Journal of Machine Learning Research*. (2004) 975–1005.
- [29] Y. Wada, M. Trincavelli, Y. Fukazawa, H. Ishida, Collecting a database for studying gas distribution mapping and gas source localization with mobile robots, *Proceedings of International Confederation of Architectural Museums (IACM)*. (2010).
- [30] S. Lloyd, Least squares quantization in PCM, *IEEE Transactions on Information Theory*. 28 (1982) 129–137. doi:10.1109/tit.1982.1056489.
- [31] G. McLachlan, D. Peel, *Finite Mixture Models*, *Wiley Series in Probability and Statistics*. (2000). doi:10.1002/0471721182.
- [32] V. Bennetts, E. Schaffernicht, V. Pomareda, A. Lilienthal, S. Marco, M. Trincavelli, Combining non selective gas sensors on a mobile robot for identification and mapping of multiple chemical compounds, *Sensors*. 14 (2014) 17331–17352. doi:10.3390/s140917331.
- [33] J. Ghosh, A. Acharya, *Cluster ensembles*, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 1 (2011) 305–315. doi:10.1002/widm.32.