**Advanced Statistical Metrics for Gas Identification System with Quantification Feedback**

<table>
<thead>
<tr>
<th>Journal:</th>
<th>IEEE Sensors Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Regular Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Brahim Belhaouari, Samir; Alfaisal University, Hassan, Muhammad; Hong Kong University Of Science and Technogy, Electronics and Computer Engineering Walter, Nicolas Bermak, A.; Hong Kong University of Science and Technology, Electronics and Computer Engineering</td>
</tr>
<tr>
<td>Keywords:</td>
<td>SYST</td>
</tr>
</tbody>
</table>
Advanced Statistical Metrics for Gas Identification System with Quantification Feedback

Samir Brahimi-Belhaouari, Muhammad Hassan, Nicolas Walter and Amine Bermak, Fellow, IEEE

Abstract—The pattern recognition problem for real life applications of gas identification is challenging due to the limited amount of data existing and the sequential variability of the mechanism mostly caused by drift and the real-time detection. These problems are commonly caused by the slow response of most of the gas sensors. In this paper, a novel gas identification approach based on the Cluster-k-Nearest Neighbour (C-k-NN) is introduced. The effectiveness of this approach has been successfully demonstrated on the experimental dataset obtained from array of gas sensors. Our classification takes advantages of both the k-Nearest Neighbour (k-NN) which is highly accurate and the K-means cluster which is able to reduce the classification time. In order to increase the accuracy rate, a new feature selection method is proposed. The selection of features is based on their ability to separate and distinguish between different classes. Advanced statistical metrics are introduced to quantify the classification contribution of each feature. Mostly, classifiers are suffering from misclassification detection; new statistical metrics are introduced to estimate the exactness of the classifier response, i.e. to detect the misclassification. To enhance the classification performances for gas identification, a new Tree classification design is introduced, named Tree C-k-NN. In order to assess the technique, experiments were conducted on six different gases. Accuracy rate of 98.7% have been obtained with the C-k-NN and 100% with the Tree C-k-NN. The performance of this approach is also validated by using three publicly available data sets.

Index Terms—Clustering, feature selection, K-Means, Gaussian Mixture Model (GMM), classification, gas identification, confidence coefficient.

I. INTRODUCTION

REAL-TIME applications are demonstrated as a critical issue in a wide range of applications especially in the civil and military environments [1]. In gas identification systems, metal oxide semiconductor gas sensors are regularly preferred for the merit of high selectivity, small cost, modest operation circuit etc. But the noticeable disadvantages such as absence of stability and lack of generalization (one gas dedicated sensor) impede their wide application in the industrial field. Current studies often use the Temperature Modulation Dynamic Test (TMDT) method to improve their selectivity and stability; moreover, a sinusoidal signal is often used to implement temperature modulation [2], [3]. Under temperature modulation, the sensor’s dynamic response curves show different patterns when contacting different gases, and the character variables used for pattern recognition could be extracted from the sensor’s response curves [4].

On the other hand, a gas sensor array permits the selectivity of a single gas sensor and has the ability to classify different odours. In fact, an array of different gas sensors may generate a unique signature for each odour. After going through the pre-processing stage, the resulting feature vector is used to solve a given classification problem, which consists of identifying an unknown sample from a set of previously learned gases. Significant work has been devoted to design a successful pattern analysis system for machine olfaction [5], [6], [7], [8], [9], [4], [2], [10], [11], [12].

Pattern recognition algorithms have become a critical element in the effective implementation of chemical sensor arrays and electronic noses. Traditional chemical sensing techniques have relied on the inherent selectivity of the sensor to determine the presence or absence of the target analyte(s) [5].

Probabilistic Neural Networks (PNN), Learning Vector Quantization (LVQ) Neural Networks, Back-Propagation Artificial Neural Networks (BP-ANN), Soft Independent Modelling of Class Analogy (SIMCA), Bayesian Linear Discriminant Analysis (BLDA), Mahalanobis Linear Discriminant Analysis (MLDA), Probabilistic Principal Component Analysis (PPCA) and the Nearest-Neighbour (NN) pattern recognition algorithms have been developed for classifying chemical sensor data [9], [8], [2], [10]. The most notable is the Neural Network algorithm, which is capable of using the Multi-Layer Perceptrons (MLPs) and Radial Basis Functions (RBFs) for the pattern recognition [7], [2]. Other methods that are based on the class-conditional density estimation have also been used; they are the quadratic- and k-Nearest Neighbour (k-NN) classifiers. These parametric and nonparametric density estimation methods have their merits and limitations. As chemical sensor arrays are expected to operate in different types of environments and situations, the most suitable pattern recognition algorithm for this kind of application requires at least the following qualities:

- High accuracy,
- Minimum classification time,
- Simple to train,
In this paper, we take advantage of the k-Nearest Neighbour, k-means cluster and Gaussian Mixture Model to design a hybrid classifier which will correspond to these requirements, named Cluster-k-NN (C-k-NN).

As the array sensor outputs several features, it is of common use to perform feature selection [13]. Indeed, the classifier can be affected negatively due to the extraction of a too large amount of features or non-appropriate features [14], [15], [16]. Hence, Feature selection is a crucial step for pre-classification as suitable features can enhance significantly the accuracy rate. Moreover, feature selection can be seen as dimension reduction technique and consequently will reduce the computation time. For classification, the selection of the best features is equivalent to identify the most discriminant features which will distinct between different classes. For that purpose, we need to quantify the contribution of each feature to the accuracy rate of classification, i.e. the variability of the features with respect to the classes. Section II will detail two statistical metrics to locate the $n^{th}$ adequate features.

Classification using the Gaussian Mixture Model (GMM) [3] and k-Nearest Neighbour (k-NN) [4] are equivalent in the sense that they both consider the neighbour data of a new vector $x$ to be classified and are more accurate compared to NN (Nearest Neighbour). Indeed, GMM and k-NN are more efficient as they well behave with the less frequent samples of training data in the overlapping areas [17], [18].

One way to reduce the classification time of a k-NN classifier is to cluster the space of training data into subclasses [19], [20]. Each subclass contains a random number of data, which are relatively close to each other and will be represented by one or more statistical characteristics such as mean, median, variance, etc. Thus, the NN (or k-NN) algorithm will use the representative(s) of the data rather than the whole data set for classification. This kind of classification is called the Cluster-k-NN (C-k-NN), which is similar to the 'variable $k$'-NN [11], [21].

The complexity in NN and k-NN algorithms are of the order of $O(N^2)$, where $N$ is the size of the training data. The advantage of data clustering is to reduce the complexity by summarizing the data of each subclass by their representatives, low memory requirements, robust to outliers, produce a measure of uncertainty.

Thus, the complexity of C-k-NN is of the order of $o(N)$. Indeed, the number of subclasses is independent of the data size $N$. As a consequence, the proposed clustering method will behave as a GMM that has the ability to approximate an arbitrary statistical data distribution by a limited set of Gaussian functions and will scale very closely to the data without any parameter estimation related to GMM [17].

To estimate the number of subclasses and their representatives for C-k-NN, K-Means cluster may be used [22], [23], [24], [25], [10]. Although the K-Means algorithm is efficient, it requires the number of subclasses and the initialization vectors as input. The initialization vectors can be computed by Hierarchical Near-to-Near Algorithm [11], [21]. A popular heuristic for K-Means clustering is Lloyd’s [26] (1982) algorithm [27]. Several techniques have been proposed to optimally initialize the K-Means clustering based on Genetic Algorithm, Trees, statistical information, etc. [28], [29], [30], [31], [32], [33], [34], [35], [36].

In this paper, the number of subclasses (i.e. number of clusters for each class) will be determined automatically by an iterative process starting from one [11], [21]. To ensure the separation between classes two Stopping Criteria (SC) are used and detailed in Section III. The first SC maintains a maximum distance between the classes and the second SC find the optimal subclasses number.
One of the missing properties of classification systems is the measurement of the confidence of their responses. This kind of metric can be used to detect the misclassifications. Two metrics are introduced to quantify the confidence of the response of the designed classifier.

For validation purposes, real data collected from gas sensor arrays has been used. Six gases have been tested: CH4, CH4CO, CO, C2H6OCH4, C2H6OOC and C2H6O. Respectively from the gas sensor response, 25, 106, 50, 90, 111 and 50 samples have been extracted to form the dataset [8], [11], [21]. We also used breast cancer, iris and seeds data sets from [37] in order to validate the performance of our proposed algorithms in the applications from other fields.

The reminder of this paper is organized as follows, Section II defines two metrics for feature selection that can be used accordingly to the complexity of the data to be classified. Section III focuses on the classification method by describing in details the selection of the representative of the data, based on GMM and the clustering method, based on K-Means, with initialisation concern. A new quantification metric of classifier responses is introduced in Section IV to evaluate the quality of classifiers and to estimate the exactness of the classification. Section V discusses the experiments and results obtained from array of gas sensors and data sets from other applications. Finally, Section VI summarizes the research work and presents future work.

II. FEATURE SELECTION

In order to reduce the data dimension after feature extraction, a feature selection is performed. The goal is to choose the most suitable features for classification, i.e. the combination of features that will best discriminate the different classes. For that purpose, the contribution of each feature with respect to the classification needs to be quantified. Let suppose two classes l and k and feature j. The contribution of the jth feature is quantified according to the following statistical metric:

\[
\text{Metric}_j(l, k) = \frac{|\bar{x}_j(l) - \bar{x}_j(k)|}{\sqrt{\frac{S_j(l)^2}{n_l} + \frac{S_j(k)^2}{n_k}}}, \tag{1}
\]

where \(\bar{x}_j(l)\) and \(\bar{x}_j(k)\) are the average of the different values of the feature j for the class l and class k, respectively. \(S_j(l)^2\) and \(S_j(k)^2\) are the variance of the different values of the feature j for the class l and class k, respectively. \(n_l\) and \(n_k\) are the sample number contained in class l and class k, respectively.

If \(n_l\) and \(n_k\) are enough bigger, we can approximate the Probability Density Function of the defined Metric\(_j(l, k)\) by a student distribution, named \(t_{DF}\), with a degree of freedom defined as follows:

\[
DF = \frac{(n_l - 1)(n_k - 1)}{(n_l - 1)(1 - V)^2 + (n_k - 1)V^2}
\]

where

\[
V = \frac{\frac{S_j(l)^2}{n_l}}{\frac{S_j(l)^2}{n_l} + \frac{S_j(k)^2}{n_k}}.
\]

Also we can define another metric that can quantify the classification contribution of each features between two classes,
for example \( l \) and \( k \):

\[
Metric_j^2(l, k) = 1 - A_{\text{overlap}}(l, k)
\]  

(2)

Where \( A_{\text{overlap}}(l, k) \) is the overlapping area as exhibited in Fig. 1.

According to the previous graph, the area of overlapping can be defined by the sum of the minimum:

\[
A_{\text{overlap}}(l, k) = \int_{-\infty}^{+\infty} \min(\phi_l(x), \phi_k(x)) dx
\]

\[
= \min(\Phi(\frac{x_l - \mu_l}{\sigma_l}); \Phi(\frac{x_l - \mu_k}{\sigma_k}))
\]

\[
+ \min(1 - \Phi(\frac{x_k - \mu_l}{\sigma_l}); 1 - \Phi(\frac{x_k - \mu_k}{\sigma_k}))
\]

\[
+ \min(\Phi(\frac{x_y - \mu_l}{\sigma_l}); \Phi(\frac{x_y - \mu_k}{\sigma_k}); \Phi(\frac{x_l - \mu_k}{\sigma_k}))
\]

The smaller the overlapping area is, the more suitable the feature to distinguish between the two classes.

To generalize this metric to a large number of classes, we use the following formulation:

\[
Metric_j^p = \min_{l \neq k}(Metric_j^p(l, k)), \ p = 1 \ or \ 2
\]  

(3)

Fig. 2 exhibits an example of three classes well separated by two features \( x \) and \( y \). In this case, it is preferable to use the formula (3) to detect those features.

For difficult cases, the previous metric might not be able to identify any features suitable for the separation of all the classes as shown in Fig. 3. Therefore, it is possible to detect features that separate between at least one class from the others. The following metric can be used to overcome this kind of problems:

\[
Metric_j^p = \max_{l \neq k}(Metric_j^p(l, k)), \ p = 1 \ or \ 2
\]  

(4)

### III. Classification Process

#### A. Cluster-k-NN Classifier

The most widely non-parametric method for density estimation is the \( k \)-NN [24], [38]. The \( k \)-NN rule is a powerful technique used to generate a highly nonlinear classification with limited data, i.e. small training data size. To classify a pattern, we need to find the closest \( k \) samples in the dataset and select the predominant class among those \( k \) neighbours to be assigned to the tested pattern. As problem could arise if two or more classes are predominant, it is preferable to choose an odd \( k \) value. One drawback of \( k \)-NN is that the training data must be pre-stored, and a large amount of processing power is needed to classify the new pattern. However, our hybrid C-\( k \)-NN will manage to overcome these drawbacks by simplifying the density estimation.

Basically, a \( k \)-NN classifier is based on the Euclidean distance between a test sample \( x \) and the specified trained samples. However, this paper introduces a new metric in order to reduce the time consumed by the estimation of the probability density function of classes. Let \( x_{i,j} \) belong to the subclass \( j \) of the class \( i \), denoted by \( C_{i,j} \), and for all positive numbers \( s \), we may define the following metric:

\[
d(x, C_{i,j}) = \frac{d_{\text{Euclid}}(x, \hat{x}_{i,j})}{n_{i,j}}, \ \forall x \in \mathbb{R}^d,
\]

(5)

where \( \hat{x}_{i,j} \) represents the subclass \( C_{i,j} \) and \( n_{i,j} \) can be defined as the number of samples of the subclass \( j \) (\( \text{card}(C_{i,j}) \)) or either as the variance of the set \( C_{i,j} \).

One of the most commonly used parametric method is the Gaussian Mixture Model (GMM) [39], [3], [9]. GMM is classified as semi-parametric density estimation since it defines
a very general class of functional forms for the density model. In a mixture model, a probability density function is expressed as a linear combination of basic functions. A model with $M$ components is described as a mixture distribution:

$$P(x) = \sum_{j=1}^{M} P(j) P(x \mid j),$$

where $P(j)$ are the mixing coefficients and $P(x \mid j)$ are the component density functions. Each mixture component is defined by a Gaussian parametric distribution in $d$ dimensional space:

$$P(x \mid j) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_j)}} \exp\left(-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)\right),$$

where $\Sigma_j$ is the covariance matrix of the random variable $x$ giving that it belongs to class $j$. The parameters to be estimated are the mixing coefficients $P(j)$, the covariance matrix $\Sigma_j$ and the mean vector $\mu_j$ which are time consuming. To fasten the computation, the proposed C-k-NN algorithm approximates each Gaussian function, $P(j)P(x \mid j)$, which can be seen as one particular subclass, by the following formula:

$$P(j)P(x \mid j) \approx \frac{1}{\varepsilon + d(x, C_j)}.$$

where $\varepsilon$ is any small number added to avoid the division by 0 and $C_j$ is the class $j$ with mean $\mu_j$.

In this paper, a modified K-Means algorithm is proposed to cluster each class. The modified K-Means differs from the classic one by setting precisely the initialization vectors and finding the optimal number of subclasses for each class. The classic algorithm of the K-Means Cluster is not stable because the result depends on the random choice of the $k$ initial vectors. Therefore, the Hierarchical Near-to-Near and Hierarchical Near-to-Mean techniques has been developed to supervise the initialization of the modified K-Means algorithm, which in general gives better results than the classic K-Means cluster [11], [21]. Indeed, inside each class, the sum of the variances of all the subclasses is better minimized resulting in:

$$\text{var}(C_i,k)_{\text{modified}} \leq \text{var}(C_i,k)_{\text{classic}}$$

As shown in Fig. 4, the Hierarchical Near-to-Near algorithm appears more adapted to the identification of gases.

To find the number of clusters (subclasses), the number of clusters is iterated starting from one until the following Stopping Criteria (SC) are verified at each iteration:

1) All the representatives or centroids ($\mu_{i,j}$) have to be closer to their classes $C_i$ than to any other classes. This is to reduce the misclassification, i.e. reduce the classification errors of our training data.

2) The sum of the variances of the $n$ clusters, $\text{var}(C_i,n)$, does not decrease drastically in comparison to the previous iteration ($n - 1$). The variance of each class $C_i$ is defined as follows $\text{var}(C_i,n) = \sum_{j=1}^{n} \text{var}_j$, where $\text{var}_j$ is the variance of the subclasses $j$. To define the smooth function, the criterion $(\text{var}(C_i,n) - \text{var}(C_i,n-1)) \leq \alpha$ is proposed. In our simulations $\alpha = 0.9$ gives the best result. Fig. 5 and 6 illustrates the behavior of the variance criterion for two different classes.

The previous iteration process defines a dictionary referencing the generated subclasses with the classes. Each class $C_i$ is
represented by \( \{\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,m_i}\} \) with \( 1 \leq i \leq cn \), \( cn \) the number of classes and \( m_i \) the number of subclasses for the class \( C_i \).

To classify a new pattern \( x \), the NN algorithm or \( k \)-NN algorithm (with a small \( k \)) is used on the dataset \( \{\mu_{i,j}; 1 \leq i \leq cn \text{ and } 1 \leq j \leq m_i\} \). The classification rule is based on the minimum distance to the dictionary, i.e. the distance to the representative of the subclasses, \( \mu_{i,j} \).

The class \( C_i \) is assigned to the pattern \( x \) if the following formula is verified:

\[
C_i = \underset{1 \leq i \leq cn}{\text{arg min}} \{d(x, \mu_{i,j})\}, \quad 1 \leq j \leq m_i
\]

where \( \text{arg } d(x, \mu_{i,j}) = C_i \) for all \( 1 \leq i \leq cn \).

**B. Tree Cluster-\(k\)-NN classifier**

One of the ideas to increase the accuracy rate is to reduce the number of classes. Indeed, the accuracy is strongly related to the number of classes. The reduction of the number of classes can be achieved only by hierarchical classification. This paper presents a novel approach named tree Cluster-\(k\)-NN. The tree consists in the integration of the Cluster-\(k\)-NN classifier in each node to classify two groups. To design the desired tree, the classes are gathered with respect to their similarity as shown in Fig. 7.

**IV. Quantification Metric**

The error of classification inherent to all classifier must be avoided in some authentication systems. Indeed, the consequences of misclassification in these specific applications can cause important damage. Therefore, a new approach for estimating the exactness of the classification response is required to detect the misclassification. In case of misclassification, the system will re-acquire the pattern to be classified. The classification of the pattern will be performed by the same classifier or even by another classification algorithm. To quantify the classifier response, we propose to define a new statistical metric that have the tightest function probability of error, i.e. the tightness of the probability function will inform about the confidence of the classification results. Two new statistical metrics, named Confidence Coefficient (CC), are proposed based on the dictionary previously defined, \( \{\mu_{i,j}; 1 \leq i \leq cn \text{ and } 1 \leq j \leq m_i\} \). By analogy to the computation of parallel resistances in electric circuit, the first CC can be computed as follows:

\[
CC_1 = 1 - \frac{1}{\sum_{k=1}^{n} \frac{1}{d_i}}
\]

where \( d_1 < d_2 < \ldots < d_i \) and \( d_3, d_2, \ldots, d_i \) are the smallest Euclidean distances between the tested input datum and the nearest class and \( d_i+1 \) is the first smallest Euclidean distance to another class. Fig. 8 presents two examples that can be faced during classification. The first one gives high level of confidence and the second lower level of confidence. In some cases with complex data, the number of subclasses...
the sensor responses is recommended due to the short intervals between consecutive sampling. The second step is dedicated to the cleaning of the gas chamber with dry air for a period of 40 minutes. This longer period of time will ensure the complete removal of the gas tested before.

For validation purposes, six gases were selected for testing: CH4, CH4CO, CO, C2H6OCH4, C2H6OCO and C2H6O. The dataset contains 25, 106, 50, 90, 111 and 50 samples, respectively from the gas sensor response [22], [40]. The characteristics of the datasets are summarized in Table I.

We also evaluate the performance of our algorithms by choosing three data sets: breast cancer, iris and seeds from publicly available data records [37]. For the breast cancer data set, 30 features are extracted from digital images of tumor cells for every patient and a prediction is built on these features. The Iris data base consists data of 3 iris plants and each iris plant is characterized by 4 features. The seeds data set consists of 3 wheat varieties and every variety is characterized by 7 geometric features.

**Feature Selection Results**

The features are ranked according to the metrics described in Section II. Let’s study the influence of the best features, ranked according to the metric value, in terms of classification accuracy. Fig. 9 (a) exhibits the evolution of the classification rate in function of the number of best features used by the classifier.

As the graph is quasi monotonically increasing, it shows that the **Metric**1, defined by formula (3), ranked the features according to the importance of their contributions to the classification.

As a comparison, the **Metric**2, defined by equation (3), gives better ranking of the features as the classifier achieves higher accuracy with less features than the previous metric, i.e. Evolution of the classification rate in function of the number of best feature according to the **Metric**3, as shown on Fig. 9 (b).
Figure 18. Graph of the confidence response value on the gas sensors testing dataset with $CC_1$ and all the features (a) and the three best features (b) selected by formula (2). The red square (■) locates the misclassified data.

Figure 19. Graph of the confidence response value on the gas sensors testing dataset with $CC_2$ and all features (a) and the three best features (b) selected by formula (2). The red square (■) locates the misclassified data.

Figure 20. Graph of the confidence response values on the gas sensors testing dataset with $CC_2$ for two different sets containing all available features (a and b). The red square (■) locates the misclassified data. The occurrence of $CC$ values inferior to 0.3 is high, means that the classifier is not adequate. The clustering must be refined by using one of the metric defined in Section III.

Table V

<table>
<thead>
<tr>
<th></th>
<th>Tree Classification Results for Seed Data Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch 1</td>
<td>Classes 1 and 2 Vs Class 3</td>
</tr>
<tr>
<td>Branch 2</td>
<td>Class 1 Vs Class 2</td>
</tr>
</tbody>
</table>

As the Metric$^2$, defined in formula (3), reach a high accuracy, means the classes are well separated by the features selected, there is no necessity to go further to Metric$^2$, defined by formula (4). Fig. 10 illustrates alternative results obtained with this metric. As expected, the results are worse due to the high disorder in the ranking.

The effectiveness of feature ranking approach towards classification can also be observed in Fig. 11, Fig. 12, Fig. 13, Fig. 14 and Fig. 15 for breast cancer, iris and seed data sets.

Classification Results

The test results, summarized in Table II, show that the proposed approach increase significantly the accuracy rate with a limited classification time. Moreover, the cross validation
shows the non-dependency of the accuracy with respect to the training data.

According to the sensitivity of the application that might require a perfect classifier, i.e., no tolerance to the misclassification, the Tree Cluster-k-NN classifier is used to achieve an accuracy rate of 100%. On each node, our classification algorithm is applied to separate one gas from the remaining gases. Fig. 16 illustrates the classifier designed to classify the six gases without misclassification.

Table III resumes identification performances of different techniques. From the Table III, the classification by the proposed methods gives higher classification accuracy compared to previous techniques [8].

Tree Cluster-k-NN is not applicable for breast cancer data set because this data set contains only 2 classes. Table IV and Table V show classification performance for iris and seed data sets with our proposed classifier.

Confidence Coefficient Results

The estimation of the relation function between the probability of misclassification and Confidence Coefficient (CC) values can be obtained by collecting CC values from various configurations of learning and testing dataset used by the proposed classifier C-k-NN. Fig. 17 exhibits a tight function which implies that the confidence coefficient is effective, i.e., the more tight the function, the more efficient the CC for detecting misclassifications.

In case the CC value is less than 0.1, the classification response is 50% true or false, i.e., it is better to not considerate this particular response and acquire a new sample data to confirm the classification. One may consider another method of classification. In contrast the closer the CC value to one, the more confident the classification response. From Fig. 18 (a) and Fig. 19 (a), it is evident that our two metrics are very efficient in classification correctness, since almost all CC values are closed to one. 98.75% of the values are greater than 0.5. More than that, the classifier can be evaluated as not adequate, if the occurrence of a CC value less than 0.1 is high. Fig. 19 exhibits better results as the CC values are closer to one compared to CC1 values in Fig. 18.

Fig. 18 (b) and Fig. 19 (b) show the CC1 and CC2 of classification results by using the three best features selected.
by formula (2). Some classification errors could be corrected if all the results with $CC$ values less than a determined threshold (0.3) are subjected to further analysis, i.e. reacquisition request, classification by another algorithm, etc.

To highlight the interest of the defined confidence coefficient, Fig. 20 shows the classification results of two different datasets with non-optimized alpha parameter, i.e. non-optimized number of sub-classes.

Fig. 21, Fig. 22 and Fig. 23 show the CC1 and CC2 values of data samples for breast cancer, iris and seeds data sets respectively. Classification performance of the data sets could be improved by increasing the confidence values.

VI. Conclusion

In this paper, classification system using Cluster-$k$-NN is introduced. It is shown that the designed classifier is very powerful in term of accuracy and efficient in term of computational time. It inherited the advantages of the $k$-Nearest Neighbour and the Gaussian Mixture Model for the accuracy as well as the $K$-Means for the clustering to reduce the classification time.

In order to overcome the general problems encountered by classical algorithms, several measures have been introduced for feature selection, clustering and confidence level of classifier response purposes.

The proposed feature selection metrics significantly improve the performances of the system as high accuracy have been achieved with few features.

The space clustering is enhanced by the modified $K$-Means clustering and the defined two Stopping Criteria to achieve better segregation between the different classes based on their statistical behaviour.

New metrics have been proposed to quantify the classification response. First, to evaluate the global classification performances of the classifier, i.e. the classifier responses can be trusted. Secondly, some classification errors could be corrected.

The C-$k$-NN algorithm has been tested on a dataset composed of six gases and achieves 98.7% accuracy. The performance of our proposed algorithms is also validated on publicly available data sets. A dedicated Tree C-$k$-NN has been designed to reach the need of some applications that do not tolerate errors of classification, i.e. 100% accuracy. The Tree C-$k$-NN is suitable to face the problematic of a complete gas identification system, i.e. complete gas dictionary.

Future works will be oriented on the refinement of the misclassification detection by improving the introduced metrics with statistical measures.

ACKNOWLEDGMENT

The authors are grateful and indebted to the support given by all individuals in Universiti Teknologi PETRONAS for this work and the Hong Kong University for sharing their Gas database.

REFERENCES


Samir Brahim-Belhouari received the Master’s degree in Telecommunications in 2000 from Institut Nationale Polytechnique (INP) of Toulouse, France and the PhD in 2006 from the Federal Polytechnic School of Lausanne (EPFL). He is currently an Assistant Professor with the Math and Computer Science Department of Alfaaisal University, Kingdom of Saudi Arabia. His research interests lie in the areas of image and speech processing and stochastic processes.

Muhammad Hassan received his MPhil degree in Electronics and Computer Engineering from Hong Kong University of Science and Technology (HKUST) in 2012. He is currently pursuing for the PhD degree in the same university. His interests include development of pattern recognition algorithms for odor discrimination and development of energy aware platform for structural health monitoring applications.

Nicolas Walter received the master’s degree in electrical engineering with instrumentation and image processing specialties from the Université de Bourgogne in 2006 and the PhD degree in 2010 from the same university at the 3D cell of the L2EI, France. Currently, he is Post-doctoral researcher in Universiti Teknologi PETRONAS, Malaysia. He is involved in different projects for industry and biomedical applications. His research interests include computer graphics, 2D and 3D edge/corner detection, multi-scale analysis, 3D reconstruction, multi-spectral analysis and classification.
Amine Bermak (M’99-SM’04-FM’13) received both his MEng and PhD degrees in Electronic Engineering from Paul Sabatier University, Toulouse, France in 1994 and 1998, respectively. During his PhD, Dr. Bermak was part of the Microsystems and Microstructures Research Group at the French National Research Center LAAS-CNRS where he developed a 3D VLSI chip for artificial neural network classification and detection applications.

He then joined the Advanced Computer Architecture research group at York University, York, England, where he was working as a Post-doc on VLSI implementation of CMM neural network for vision applications in a project funded by British Aerospace. In 1998, Dr. Bermak joined Edith Cowan University, Perth, Australia, first as a research fellow working on smart vision sensors, then as a lecturer and a senior lecturer in the school of Engineering and Mathematics. He is currently a Professor with the Electronic and Computer Engineering Department of Hong Kong University of Science and Technology (HKUST) where he is also serving as the director of Computer Engineering and the director of MSc degree in Integrated Circuit Design (ICDE).

Dr. Bermak has received many distinguished awards, including the 2004 "IEEE Chester Sall Award"; "Michael G. Gale Medal", 2011; and the "Best Paper Award" at the 2005 International Workshop on System-On-Chip for Real-Time Applications. Dr. Bermak is a member of technical program committees of a number of international conferences including the IEEE Custom Integrated Circuit Conference CICC’2006, CICC’2007, the IEEE Consumer Electronics Conference CEC’2007, Design Automation and Test in Europe DATE2007, DATE2008. He is the general co-chair of the IEEE International Symposium on electronic design test and applications, Hong Kong 2008, and the general co-chair of the IEEE Conference on Biomedical Circuits and Systems, Beijing, 2009. He is also on the editorial board of IEEE Transactions on VLSI systems; IEEE Transactions on Biomedical Circuits and Systems and the Journal of Sensors. Dr Bermak is a fellow member of IEEE and a member of IEEE CAS committee on sensory systems. His research interests are related to VLSI circuits and systems for signal, image processing, sensors and microsystems applications. He has published extensively on the above topics in various journals, book chapters and refereed international conferences.