Sensor Array Optimization for Complexity Reduction in Electronic Nose System

Md. Mizanur Rahman [*1, ***], **Chalie Charoenlarpnopparut** [**2], **Prapun Suksompong** [*3], and **Pisanu Toochinda** [*4], Non-members

ABSTRACT

An Electronic nose (E-Nose) can be used to assess food quality and fruit ripeness without personal bias. A set of relevant sensors must be identified to design an effective E-Nose and reduce implementation cost and complexity. The analysis of tropical fruit odour in terms of pattern recognition errors is carried out to determine the minimum number of sensors and their combinations. Two new methods namely 1) principal component loading and mutual information between sensor data, and 2) threshold based approach are proposed in this work to evaluate and optimize the sensor set. Four pattern recognition methods, namely multilayer perceptron neural network (MLPNN), radial basis function neural network (RBFNN), support vector machine (SVM), and k-nearest neighbour (k-NN) are also compared in terms of classification performance. The pattern recognition error of SVM with the optimal set of sensors is as low as 2.78% and that of k-NN is 9.72%. The results conclude that the pattern classification error with MLPNN, and RBFNN is higher than the error from k-NN and SVM.

Keywords: Electronic Nose, PCA, MLP, RBF, SVM, k-NN, Sensor Optimization.

1. INTRODUCTION

An electronic nose (E-Nose) primarily is a combination of an array of gas sensors with diverse selectivity, a data acquisition system, and a pattern recognition system. An E-Nose [1-2] mimics a mammalian olfactory system [3]. It is being applied to environmental monitoring, medicine, food industry, homeland security, personnel security, product-development research, etc. [4-5]. Large scale implementation of E-Nose requires reduction of manufacturing cost and computational complexity. Reducing the number of sensors decreases manufacturing cost as well as computational complexity which arise due to dimensionality. In order to reduce the design and algorithm complexity, the sensors of an E-Nose sensor panel with insignificant information are discarded, and more significant ones are chosen. The optimal set of sensors is chosen by different sensor array optimization techniques proposed in the literature [6-9, 11-12].

A sensor sub-array based method is applied to reduce the number of sensors [6] to classify eleven gases. Firstly, sub-arrays are formed by the selectivity differences of each sensor. Later, sensors from different sub-arrays are combined to obtain the sensor array with minimum classification error. To discriminate different grades of Longjing tea analysis of variance (ANOVA) and principal component (PC) loading are combined to reduce the number of sensors [7]. The PC loading method is also used in [8] to exclude redundant sensors. Here it was shown that ANOVA and Tukey-multiple-comparison suggest similar choice of the sensors. Tukey-multiple-comparison compares the distances between means of each sensor data with a constraint that all the classes lie in a single group. Both ANOVA and Tukey-multiple-comparison do not provide individual sensorâĂŹs capability to classify classes. In addition, methods presented in [7-8] do not show any policy where third or higher dimensional PCs may contain significant data variances.

A genetic algorithm (GA) is used in [9] to optimize an E-Nose sensor array and determine the tea quality. For GA an individual gene is represented by a sensor. Genes are combined to construct chromosomes. A number of such chromosomes (combinations of sensors) form a population. GA technique aims to reduce the number of sensors (gene) in the combination which limit the length of the constructed chromosome. The fixed chromosome length overlooks the longer or shorter chromosomes which can mislead to suboptimal set of sensors. To overcome this problem, one way is to perform exhaustive search into all combinations of sensors and apply a classification algorithm to find the classification errors. The optimal sensor set is the one which meets the acceptable pattern recognition errors (set by the designer or application) with minimum number of sensors. In addition,

Manuscript received on September 28, 2016 ; revised on October 21, 2016.

* The authors are with School of Information Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand, E-mail : mizan.ku@gmail.com[1], E-mail : chalie@siit.tu.ac.th[2], E-mail : First.prapun@siit.tu.ac.th[3]

** The authors are with School of Bio-chemical Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand, E-mail : pisanu@siit.tu.ac.th[4]

*** The authors are with Electronics and Communication Engineering Discipline, Science Engineering and Technology School, Khulna University, Bangladesh

global optimum cannot be achieved all the time with GA, especially when overall solution has various populations. The GA is more fitted for the situations that tolerate trial and failure results [10].

In order to design a low cost E-Nose, the number of sensors and total complexity of analysis must be minimized. The exhaustive search and GA, both the methods, search for the optimal set of sensors out of many combinations of sensors. The pattern recognition performance of every sensor combination is evaluated by a pattern recognition algorithm as many times as there are number of sensor combinations. This leads to high complexity. This work aims to select the proper sensors, reduce the number of sensors and diminish the complexity of analysis and cost with satisfactory results.

The proper sensors for an E-Nose are selected based on the ratio of between (i.e. variance between classes) to within variance (i.e. average of individual class variances) of the classes [11]. Higher values of this ratio were used to select the proper sensors for the E-Nose in [12] to identify ethanol, 2-propanol, acetone, and ammonia. This method assumes equal within class variances. The performance of the optimal sensor set was verified by a multilayer perceptron neural network (MLPNN). In the results and discussion section we compare the ratio of between to within variance method [11] in contrast to our proposed methods of choosing optimal sensor set by exclusion of the redundant sensors. The exhaustive search, GA, and ratio of between to within variance methods do not consider the relationships between sensor responses. Thus a possibility remains of choosing sensors with similar information content. The principal component loading & mutual information based method and threshold based methods are proposed to solve this problem.

In this research, an E-Nose sensor panel is designed with eight metal oxide gas (MOG) sensors, which is used to sense volatile organic compounds (VOCs) from four different fruits at three ripeness states. The E-Nose sensor array optimization techniques were evaluated. Two methods were proposed to reduce the number of sensors which are PC loading & mutual information approach and a threshold based approach.

In [13-14], support vector machine (SVM) and k-nearest neighbour (k-NN) is shown to have good classification and pattern recognition performance. We also apply SVM and k-NN in this work. In addition, two neural network based regression algorithms, namely MLPNN and radial basis function neural network (RBFNN) [13-14] are also applied to compare their classification performance in this application research. These four pattern recognition methods were evaluated in this study to be used as a tool for our e-nose design. Although classification and pattern recognition performance of SVM and k-NN is good, they have higher implementation complexity compared to MLPNN, and RBFNN. The SVM requires expensive quadrature programming to find the support vectors, while k-NN applies exhaustive search to find the k-nearest neighbours which is also expensive.

2. MATERIALS AND METHODS

Sample collection, designing sensor panel, experimental setup, and sensor panel optimization techniques are discussed in this section.

2.1 Sample Collection

Four kinds of fruits, namely, sapodilla, pineapple, banana, and mango, are chosen for the experiment. Change in ripeness state of these fruits is fast, which may cause losses to businesses and consumers. During an experiment with one kind of fruit, the other fruits were kept in separate boxes to isolate their chemicals/odour to prevent any interferences from other VOCs.

2.2 Experimental Setup and Procedure

The E-Nose system consists of a sample chamber, measurement chamber, and data acquisition and classification system. The airtight sample chamber and the measuring chambers are connected by two one-inch transparent plastic tubes, via unidirectional control valves (3 and 4 in Fig.1) along with their corresponding DC fans (3 and 4), to circulate the sample headspace between them during measurement. After each experiment, unidirectional control valves (1 and 2 in Fig.1) and DC fans (1 and 2 in Fig.1) are used to circulate the free air through the measuring chamber to achieve base level sensor responses. A DC power supply is used to power the fans, sensors, and data acquisition device. A double pole double throw switch is used to switch the unidirectional 1.08 watt fans (1, 2) and (3, 4) (Fig.1), alternatively. The measurement process contains two different phases: concentration, and measurement. Sample fruits are stored at 28oC throughout the experiment.

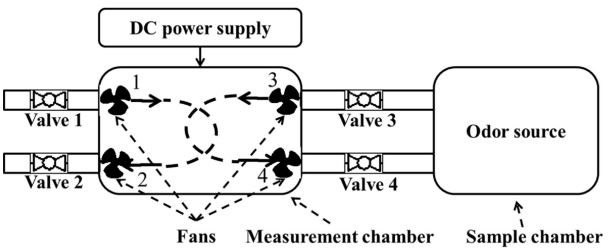

Fig.1: *The E-Nose Experimental Setup.*

The fruit samples are kept in the sample chamber. The sensor panel and data acquisition devices are kept in the measurement chamber (Fig.1). A five volt power supply is used for circuit voltage and

heater voltage of the sensors. A wireless connection is made between the data acquisition device and LabVIEW for data collection. Sensor responses to the VOCs/odours are recorded as voltage responses across the corresponding sensor loads.

We use eight metal oxide gas (MOG) sensors, listed in Table 1. The VOCs to which the sensors are responsive are also shown in Table 1. In contact with VOCs corresponding to an odour, the sensor resistance decreases which cause more current to flow through the corresponding load resistor, and thereby increases the output voltage. In the absence of VOCs, adsorption of ambient oxygen increases the sensor resistivity, and decreases the circuit current as well as the load voltage. The E-Nose sensor panel is shown in Fig.2. The panel consists, eight MOG sensors (Table 1), corresponding load resistors, two relay switches, one double pole double throw switch, and two voltage regulator ICs to stabilize circuit and heater voltages. fulfils the required pattern recognition error perfor-

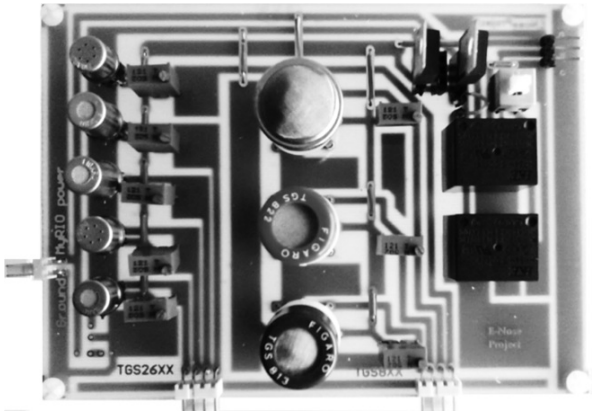

Fig.2: *E-Nose Sensor Panel.*

mance, is the optimal sensor set.

PCA loading and mutual information based approach

PCA analysis of 70% randomly chosen training data is done. The PC loading information is recorded in matrix A as in (1).

. (1)

where, $a_{i,j}$ indicates loading of sensor i on PC j . The columns of A are the PC vectors, i.e. the PCs.

2.3 Sensor Panel Optimization Techniques

In this section we discuss three methods of sensor array optimization techniques, namely, exhaustive search approach, PC loading and mutual information based approach, and threshold based approach.

Exhaustive search method

With N_s equal to eight sensors, total $2_{N_s} - 1$ (i.e. 255) sensor combinations are possible, which includes one sensor cases to eight sensor case. The pattern recognition and classification performances of all the sensor combinations are evaluated by SVM, k-NN, MLPNN, and RBFNN. The SVM and k-NN are trained with 70% randomly chosen samples from each class. The remaining 30% of the samples are used to verify the pattern recognition capability. For MLPNN and RBFNN 70% of the data are randomly chosen for training, 15% for validation, and remaining 15% for testing. The sensor combination(s) with minimum number of sensors which

The E-Nose sensor data within each class are considered to be randomly distributed with corresponding means and variances. The mutual information theory is used to measure dependency among the E-Nose sensor data. The mutual information between two random variables, and is given by,

. (2)

where, $\rho_{X,Y}$ indicates correlation coefficient between random variables X and Y [15] and is given by,$\rho^2_{X,Y} = E\left[\left(\frac{X-E(X)}{\sigma_X}\right)\left(\frac{Y-E(Y)}{\sigma_Y}\right)\right]$. Sensors having approximately similar value and direction of PC loadings give $\rho_{X,Y} > 0$ which causes large mutual information by (2). Fig.3 shows a loading plot of the experimental data on first two PCs. In Fig.3, mutual information between TGS822 and TGS813 will be large as they are similarly inclined to the PCs. Among these two sensors TGS822 should be chosen as its loading to PC1 is larger. On inclusion of every sensor to the optimal sensor cluster, the error performance of the cluster is evaluated. If the performance is not satisfactory, another sensor is added to the optimal sensor group following similar procedure. Maximum loadings on positive and negative side of each PC are identified at the beginning. First and second sensors of the optimal sensor set should be picked based on the maximum loadings on PC 1, third and fourth sensors should be picked based on the maximum loadings on PC 2, and so on. If all the loadings on any PC are unipolar, a single sensor should be picked based on that PC loadings. For example, Fig. 3 shows that all the loadings on PC 1 are positive, thus we can choose only the first sensor based on this PC loadings. In this case, second and third sensors should be picked based on loadings on PC 2. More detail is presented in results and discussion section. The mutual information of the training data between each pair of sensors is calculated by (2) and stored in the matrix B as in (3).

. (3)

where, $S_{i,j}$ is the mutual information between sensors i and j . The diagonal elements in matrix B are self-information and are not considered for the analysis.

Table 1: *Figaro Gas and VOC Sensors Used in the E-Nose Design.*

Sensor model	Sensor Index	CH_2	C_2H_2	C_3H_8	C_4H_{10}	H_2	H_2S	CO	C_6H_6	NH_3	$(CH_3)_2CO$	C_6H_{14}	Trimethyl amine and Methyl mercaptan
TGS 2612	S1	√	√	√	√								
TGS 821	S2	√	√			√		√					
TGS 822	S3	√	√		√			√	√		√	√	
TGS 813	S4	√	√	√	√	√		√					
TGS 2602	S5		√			√	√			√			
TGS 2603	S6		√			√	√						√
TGS 2620	S7	√	√		√	√		√					
TGS 2610	S8	√	√	√	√	√							

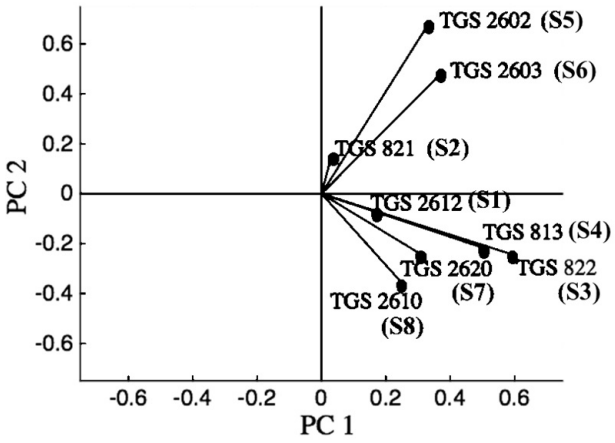

Fig.3: *PC Loading of Fruit Data to Explain the Relationship of PC Loading to Mutual Information.*

Threshold based approach

If the ratio of difference between means to summation of standard deviations of two Gaussian random variables (defined as in (4)) is large then the Gaussian plots of the variables are sufficiently apart or negligibly overlap. Fig.4 shows normalized distribution of TGS822 sensor data for three classes, namely green mango, ripe banana, and ripe sapodilla. From Fig.4 it is seen that the bell shapes of green mango and ripe banana negligibly overlap as their means are sufficiently apart compared to the summation of their standard deviations. The bell shapes of ripe banana and ripe sapodilla have large overlapping region as their means are not far compared to the summation of their standard deviations. Due to these phenomena green mango and ripe banana are easily classifiable, whereas ripe banana and ripe sapodilla might be hard to classify by TGS822. Thus a pair of classes is classifiable by a sensor if the respective Gaussian curves marginally overlap. For the mth sensor we define a ratio

$$\cdot \qquad (4)$$

where, $\mu_{i,m}$ and $\mu_{j,m}$ are the means of classes i and j, respectively; $\sigma_{i,m}$ and $\sigma_{j,m}$ are the standard deviations of the classes i and j, respectively. The larger the values of $\alpha_{i,j,m}$, lesser the overlapping of the Gaussian curves corresponding to classes i and j for sensor m. The ratio $\alpha_{i,j,m}$, for all pairs of classes are calculated for each sensor according to (4). A threshold α_{th} is set by a designer based on the expected classification error as well as the overlapping level of the Gaussian distributions of the corresponding sensor data. For any Gaussian variable, 99.7% of the data remain within three standard deviations on both sides of the mean. Thus the sensor m classifies classes i and j , if $\alpha_{i,j,m} \geq \alpha_{th} > 3$. The sensor that classifies maximum number of class pairs is chosen first, and the pattern recognition error is calculated. If the error is not within the acceptable limit as set by the designer or required by the application, a second sensor should be combined to the first one such that these two sensors together classifies higher number of class pairs. In this way more sensors are added until the required error limit is reached. The sensor combination that classifies the classes with the minimum pattern recognition error is the optimal sensor set.

2.4 Algorithms applied

Commonly used pattern recognition algorithm for E-Nose are k-NN, SVM, MLPNN, RBFNN, ordinary least squares (OLS), principal component re-

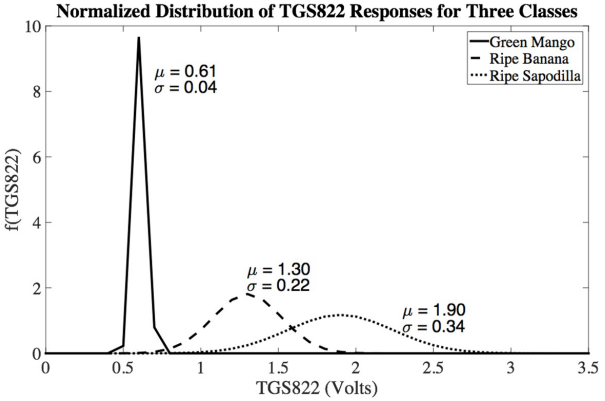

Fig.4: *Normalized Distribution of TGS822 for Three Classes with Means (μ) and Standard Deviations (σ).*

gression (PCR), and partial least squares regression (PLSR). The OLS, PCR, and PLSR are good regression techniques for pattern recognition and classification. These regression methods combine independent variables to produce a set of dependent variables. For PCR and PLSR these dependent variables are known as PCs. To reduce dimensionality the PCs which comprise most data variance are considered. Although these methods are widely used for dimensionality reduction and classification but are not suitable to find optimal set of sensors (i.e. independent variables). As this research is focused to reduce the redundant sensors (i.e. independent variables) we do not apply regression techniques such as OLS, PCR, or PLSR in this research.

In this research the pattern recognition/classification errors are analysed with four different algorithms. For k-NN the value of k is taken as 13 based on the idea that the value of k should be an odd number and should not be a multiple of the number of classes. The implemented SVM does the classification by accumulating results from multiple binary SVMs. An MLPNN is designed with 8, 10, and 1 neuron at the input, hidden, and output layers, respectively. The RBFNN is an exact RBFNN with 8, 168, and 1 neuron at the input, hidden, and output layers, respectively. Although the required number of neurons is higher in RBFNN, its training is fast compared to MLPNN.

3. RESULTS AND DISCUSSION

The voltage responses of the sensors for an experiment are shown in Fig.5. The figure shows responses of the sensor panel to ripe sapodilla VOCs. The response curves have two segments, transient and steady state. Average value of the steady state segment for every sensor is calculated. The combination of these average values forms a signature pattern for an experiment. Training data is a collection of signatures from all the experiments. The experiments for each type of the fruits at three ripeness states are repeated 20 times.

Fig.6 shows a three dimensional PCA scores plot of the training data of four fruit types (banana, mango, sapodilla, and pineapple) at three different ripeness states. The labels G, R, Rt indicate unripe, ripe, and rotten states of fruits, respectively, while B, M, S, and P represent banana, mango, sapodilla, and pineapple, respectively. (For example, RtB stands for rotten banana.)

It is seen from Fig.6 that the scores of ripe banana and ripe pineapple do not overlap with any other classes and are fully classifiable, while the scores of the other fruits at different ripeness states overlap. This overlapping indicates that the corresponding classes are not fully classifiable and some pattern recognition or classification errors are likely to occur.

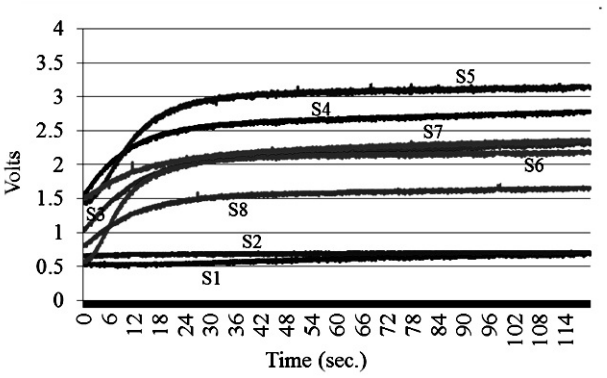

Fig.5: *Sensor Responses from Eight Sensors for Ripe Sapodilla.*

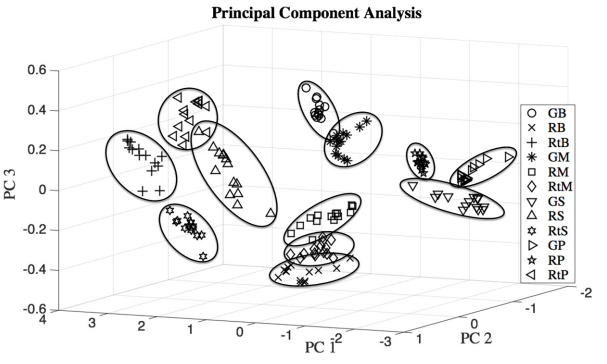

Fig.6: *PCA Scores Plot of Training Data of Four Fruit Types at Three Ripeness States.*

3.1 Between to within variance based method [11-12]

In Fig.7 the between to within class variances for each sensor are shown. To find the optimal set of sensors, the sensors are sorted as, S3, S4, S5, S6, S7, S8,

S1, and S2, as per their descending values of between to within class variances. As S3 has the highest ratio of between to within variance, it is picked first. Next, two sensors (S3 and S4) are picked, and then three sensors (S3, S4, S5) are picked, and so on. The percentage of pattern recognition error caused by, MLPNN, RBFNN, k-NN, and SVM classification algorithms for different combination of sensors is listed in Table 2. The MLPNN and RBFNN show very high classification errors compared to k-NN and SVM. We see that the pattern recognition errors with k-NN and SVM algorithms are less than 10% for three (S3, S4, S5) or more sensor combinations. Thus the three sensor combination (S3, S4, S5) can be considered as the optimal set of sensors with k-NN, or SVM chosen as classification algorithm. For improved performance sensor combinations with more sensors should be preferred, as shown in Table 2.

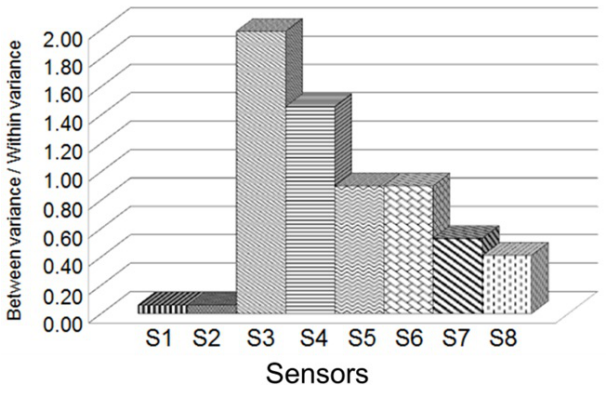

Fig.7: *Between to Within Variance of All the Classes for Each Sensor.*

3.2 Exhaustive search method

Pattern recognition capability of all the (255) combinations of sensors are evaluated by MLPNN, RBFNN, *k*-NN, and SVM. The sensor combinations with minimum pattern recognition errors are recorded in Table 4. We see in Table 3 and Table 4 that MLPNN and RBFNN show higher error compared to *k*-NN and SVM. The sensor combinations with three or more sensors show less than 10% errors for *k*-NN and SVM. For the three sensor combinations both *k*-NN and SVM show 9.72% and 2.78% errors, respectively. It is observed that the three sensor combination (S3, S5, S8) is common in both the cases. For more than three sensor cases the pattern recognition error decrease more for SVM, compared to *k*-NN.

3.3 Proposed approaches

The proposed method to reduce the number of sensors for an E-Nose is presented in this section.

Table 2: *Pattern Recognition Error Rates of Test Data for Between to Within Variance Method.*

Sensor combination with minimum error	Percent error			
	MLPNN	RBFNN	*k*-NN	SVM
(S3)	88.89	87.50	20.83	37.50
(S3, S4)	94.44	91.67	11.11	11.11
(S3, S4, S5)	80.56	69.44	6.94	4.17
(S3, S4, S5, S6)	72.22	80.56	6.94	2.78
(S3, S4, S5, S6, S7)	72.22	54.17	5.56	2.78
(S3, S4, S5, S6, S7, S8)	55.56	63.89	4.17	2.78
(S3, S4, S5, S6, S7, S8, S1)	55.56	33.33	0.00	1.39
(S3, S4, S5, S6, S7, S8, S1, S2)	19.44	27.78	0.00	0.00

PCA loading and mutual information based approach

Table 5 and Table 6 show the PC loadings and mutual information between each pair of sensors, respectively. The diagonal elements of Table 6 are self-information and are omitted as they are not required for this method. The sensor pair (S3, S7) has the largest mutual information (Table 6) and S3 has higher loading on PC1 (Table 5). Thus, sensor S3 is chosen. The sensor pair (S7, S8) has the second largest mutual information (Table 6) and S8 has higher loading on the negative PC2 axis (Table 5). Thus, S8 is chosen from the pair (S7, S8). The sensor pair (S5, S6) has the third largest mutual information (Table 6) and S5 has higher loading on the positive PC2 axis (Table 5). Thus, S5 is chosen from the pair (S5, S6). In this way, the minimal set of sensors is (S3, S5, S8), with 9.72% pattern recognition error for k-NN, and 2.78% for SVM. For the sensor combination (S3, S5, S8) the pattern recognition errors with MLPNN and RBFNN are 77.78% and 87.50%, respectively. High error with MLPNN and RBFNN indicates that they are not better choice when data overlapping exist.

Threshold based method

For any Gaussian variable, 99.7% of the data remain within three standard deviations on both sides of the mean. Thus according to equation 4, a choice for the threshold α_{th} equal to 3 will not pick any sensor for which different classes overlap. This confirms that the algorithm will pick those sensors which cause less overlapping and thereby reduce the number

Table 3: *Classification Error by MLPNN and RBFNN Methods. For Different Number of Sensor Cases, the Combinations with Minimum Pattern Recognition Errors are Recorded in the Table.*

MLPNN			RBFNN		
Sensor combination		Pattern recognition (%)	Sensor combination		Pattern recognition error (%)
Single sensor case	(S7), (S8)	80.56	Single sensor case	(S8)	79.17
Two sensor cases	(S2, S8), (S5, S8), (S6, S8)	66.67	Two sensor cases	(S4, S5), (S6, S8)	60.06
Three sensor cases	(S3, S6, S8), (S5, S6, S8)	44.44	Three sensor cases	(S2, S4, S5)	62.50
Four sensor cases	(S1, S2, S5, S8), (S2, S5, S6, S8), (S3, S5, S6, S8)	52.78	Four sensor cases	(S1, S2, S5, S7), (S1, S2, S5, S8)	41.67
Five sensor cases	(S1, S2, S4, S5, S7), (S1, S2, S4, S5, S8), (S2, S3, S5, S7, S8)	38.89	Five sensor cases	(S1, S2, S3, S4, S5), (S1, S2, S4, S5, S8)	33.33
Six sensor cases	(S1, S2, S3, S5, S6, S7)	38.89	Six sensor cases	(S1, S2, S3, S5, S6, S8), (S1, S3, S4, S5, S7, S8), (S2, S3, S5, S6, S7, S8)	31.94
Seven sensor cases	(S1, S2, S3, S4, S6, S7, S8)	27.78	Seven sensor cases	(S1, S2, S3, S4, S5, S6, S7) , (S2, S3, S4, S5, S6, S7, S8)	31.94
Eight sensor cases	(S1, S2, S3, S4, S5, S6, S7, S8)	19.44	Eight sensor cases	(S1, S2, S3, S4, S5, S6, S7, S8)	27.78

of pattern recognition errors. Smaller threshold raises error limit, and larger threshold decreases the error limit with 3 as the optimal threshold. Four kinds of fruits at three ripeness states make 12 classes. The total number class pairs are 66. Each pair is classified sequentially with each sensor and the corresponding errors are recorded. The smallest group of sensors that meets the desired error limit is chosen. The three and four sensor combinations, the number of class pairs they classify, and corresponding pattern recognition errors are listed in Table 7 and Table 8. Table 7 and Table 8 also show pattern recognition errors when 70% or all the data are applied as test data. It is seen from Table 7 that the three sensor combination, (S3, S5, S8) has the capability to classify maximum pairs of classes, compared to the other three sensor combinations irrespective of considered data volume. From Table 8, four sensor combinations, (S3, S4, S5, S8), (S3, S5, S6, S8), and (S3, S5, S7, S8) classify more pairs of classes than the other four sensor combinations for both 70% and all data cases. The classification performance of the three sensor combinations and four sensor combinations noted above are verified by MLPNN, RBFNN, k-NN and SVM. For both three and four sensor combinations MLPNN and RBFNN show large classification errors. We find again that for k-NN and SVM the three sensor combination (S3, S5, S8) has 9.72% and 2.78% pattern recognition errors, respectively. Among the four sensor cases, (S3, S4, S5, S8) shows 8.33% and 4.17% errors for k-NN and SVM, respectively. The other sensor combinations as show higher number of errors (Table 7 and Table 8). Thus with the threshold based method we find that the three sensor combination (S3, S5, S8) is the optimal sensor set.

Implementation complexity of both PC loading and mutual information, and threshold based approaches are similar. With the PC loading and mutual information based approach, and threshold based approach the classification algorithm is needed to be simulated few number of times to find the optimal sensors. The total complexity of the PC loading and mutual information based approach is the sum of the PCA complexity, complexity of finding mutual information, and the number of trials multiplied by the complexity of the classification algorithm. Whereas

Table 4: *Classification Error by k-NN and SVM Methods. For Different Number of Sensor Cases, the Combinations with Minimum Pattern Recognition Errors are Recorded in the Table.*

***k*-NN**			**SVM**		
Sensor combination		**Pattern recognition error (%)**	**Sensor combination**		**Pattern recognition error (%)**
Single sensor case	(S8)	27.78	**Single sensor case**	(S5)	30.56
Two sensor cases	(S3, S8), (S4, S6), (S5, S7), (S6, S7), (S6, S8)	15.28	**Two sensor cases**	(S5, S8)	4.17
Three sensor cases	(S3, S5, S8), (S4, S5, S7), (S4, S5, S8)	9.72	**Three sensor cases**	(S3, S5, S7), (S3, S5, S8), (S5, S6, S7)	2.78
Four sensor cases	(S3, S4, S5, S6), (S3, S4, S5, S8), (S4, S5, S6, S7), (S4, S5, S7, S8)	8.33	**Four sensor cases**	(S3, S5, S6, S7), (S3, S5, S6, S8)	1.39
Five sensor cases	(S3, S4, S5, S6, S7), (S3, S4, S5, S6, S8), (S3, S4, S5, S7, S8)	8.33	**Five sensor cases**	(S3, S4, S5, S6, S7), (S3, S4, S5, S7, S8), (S3, S5, S6, S7, S8)	2.78
Six sensor cases	S3, S4, S5, S6, S7, S8	9.72	**Six sensor cases**	S3, S4, S5, S6, S7, S8	2.78
Seven sensor cases	(S1, S2, S3, S4, S5, S6, S8), (S1, S2, S3, S4, S5, S7, S8)	8.33	**Seven sensor cases**	(S2, S3, S4, S5, S6, S7, S8)	0.00
Eight sensor cases	(S1, S2, S3, S4, S5, S6, S7, S8)	5.56	**Eight sensor cases**	(S1, S2, S3, S4, S5, S6, S7, S8)	0.00

Table 5: *Principal Component Loadings.*

Principal Component (PC) / **Sensors**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
S1	0.0954	–0.0401	0.1541	–0.2580	0.3402	0.8735	0.1418	–0.0113
S2	0.0307	0.1511	0.6316	0.4201	0.5535	–0.1618	–0.2200	0.1411
S3	0.5877	–0.2413	–0.3490	0.6504	–0.0319	0.2050	–0.0893	–0.0231
S4	0.4995	–0.2156	0.6289	–0.1967	–0.5099	–0.0467	0.0684	–0.0543
S5	0.3328	0.6816	–0.1111	–0.2193	–0.0844	0.0747	–0.5773	–0.1353
S6	0.3689	0.4790	–0.0792	–0.0222	0.1852	–0.2032	0.7426	0.0259
S7	0.2974	–0.2336	–0.1785	–0.3790	0.2305	–0.1776	–0.1599	0.7558
S8	0.2444	–0.3497	–0.0976	–0.3210	0.4716	–0.2969	–0.0930	–0.6215

Table 6: *Mutual Information Between Pairs of Sensors. The self-information in diagonal cells is omitted.*

Sensor Index	S1	S2	S3	S4	S5	S6	S7	S8
S1	-	0.0516	0.7881	0.9682	0.3070	0.4602	0.8097	0.6833
S2	0.0516	-	0.0197	0.0555	0.1382	0.1151	0.0061	0.0005
S3	0.7881	0.0197	-	1.4419	0.4015	0.6539	1.6569	1.0766
S4	0.9682	0.0555	1.4419	-	0.3783	0.6104	1.3474	1.0017
S5	0.3070	0.1382	0.4015	0.3783	-	1.5397	0.2889	0.1517
S6	0.4602	0.1151	0.6539	0.6104	1.5397	-	0.4842	0.2879
S7	0.8097	0.0061	1.6569	1.3474	0.2889	0.4842	-	1.5770
S8	0.6833	0.0005	1.0766	1.0017	0.1517	0.2879	1.5770	-

Table 7: *Maximum Number of Pairs of Classes Classifiable by Combinations of Three Sensors. Classifiable Cases are Experimented for Two Different Amount of Data: 70% of the Data and All Data.*

Three sensor combinations	Number of pairs of classes classifiable out of total 66 pairs (70% of the total data are considered)	Number of pairs of classes classifiable out of 66 pairs (all data considered)	Detection errors (%)			
			MLPNN	RBFNN	*k*-NN	SVM
S3, S4, S5	52	50	80.56	69.44	11.11	4.17
S3, S5, S7	52	50	80.56	91.67	12.50	2.78
S3, S5, S8	54	53	77.78	87.50	9.72	2.78
S5, S6, S8	53	52	44.44	98.61	13.89	4.17
S5, S7, S8	53	51	72.22	94.44	13.89	5.56

Table 8: *Maximum Number of Pairs of Classes Classifiable by Combinations of Four Sensors. Classifiable Cases are Experimented for Two Different Amount of Data: 70% of the Data and All Data.*

Four sensor combinations	Number of pairs of classes classifiable out of total 66 pairs (70% of the total data are considered)	Number of pairs of classes classifiable out of 66 pairs (all data considered)	Detection errors (%)			
			MLPNN	RBFNN	*k*-NN	SVM
S3, S4, S5, S8	55	53	72.22	48.61	8.33	4.17
S3, S4, S6, S8	54	52	63.89	62.50	12.50	6.94
S3, S5, S6, S8	55	53	52.78	61.11	12.50	1.39
S3, S5, S7, S8	55	53	72.22	77.78	12.50	5.56
S4, S5, S7, S8	54	51	63.89	54.17	8.33	6.94
S5, S6, S7, S8	54	52	63.89	88.89	12.50	2.78

with the threshold based approach some suboptimal combinations of sensors are found first, and later their pattern recognition performances are analysed with a classification algorithm. Thus complexities of the proposed methods are less, as the expensive classification algorithm is needed to be simulated only few times compared to exhaustive search or GA method. Based on the complexity and pattern recognition errors, the SVM algorithm along with the PCA and mutual information based method, or the threshold based method can be chosen to design an E-Nose sensor panel with minimum number of sensors at acceptable error rate. We find that an E-Nose can be designed picking only three sensors (S3, S5, S8) and SVM as the classification algorithm to classify banana, mango, sapodilla, and pineapple at three ripeness states with 2.78% possible pattern recognition errors. This result is also better compared to the between to within variance ratio method, where (S3, S4, S5) is the optimal sensor set with 6.94% error for *k*-NN and 4.17% error for the SVM algorithm.

4. CONCLUSION

We have presented two new approaches, one is PCA loading and mutual information based approach, and another is threshold based approach to optimize the number of sensors and analyse the pattern classification performances in contrast to, between to within variance ratio based method and exhaustive search method. It is seen that the number of sensors in an E-Nose sensor panel can be reduced with the methods proposed in this paper with low pattern classification errors. Reduction of number of sensors in an E-Nose sensor panel decreases data dimensionality as well as design cost and complexity. The classification errors with MLPNN and RBFNN are found high for this application research. The pattern recognition error performance is found better with SVM, compared to k-NN, MLPNN, and RBFNN. With the proposed sensor reduction techniques the number of sensors is reduced to three with only 2.78% classification errors for SVM and 9.72% for k-NN. It is also noted that both the proposed methods resulted same combination of optimal sensor set with k-NN and SVM pattern classification algorithms.

5. ACKNOWLEDGEMENT

The authors thank, (1) the Thailand Office of Higher Education Commission (NRU Project) and (2) Thammasat University Research Fund (Contract Number 2/22/2557) for financial support.

References

[1] K.C. Persaud, G. Dodd, "Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose," *Nature*, Vol. 299, pp.352-355, 1982.

[2] H. V. Shurmer, "An electronic nose: a sensitive and discriminating substitute for a mammalian olfactory system," *IEEE Proceedings G - Circuits, Devices and Systems*, Vol. 137, No. 3, pp. 197-204, June 1990.

[3] S. Firestein, "How the olfactory system makes sense of scents," *Nature*, Vol. 413, No. 6852, pp. 211-218, 2001.

[4] P. E. Keller, L. J. Kangas, L. H. Liden, S. Hashem, R. T. Kouzes, "Electronic noses and their applications," *Proceedings of the IEEE Technical Applications Conference (TACâĂŹ95)*, 1995.

[5] A. D. Wilson, and M. Baietto, "Applications and advances in electronic-nose technologies," *Sensors*, Vol. 9, pp.5099-5148, 2009.

[6] S. Zhang, C. Xie, D. Zeng, H. Li, Y. Liu, S. Cai, "A sensor array optimization method for electronic noses with sub-arrays," *Sensor and Actuators B: Chemical*, Vol. 142, No. 1, pp. 243-252, 2009.

[7] Z. Lei, S. Bo-lin, W. Hou-yin, and L. Zhi, "Combination Optimization Method for Screening Sensor Array of Electronic Nose," *Food Science*, Vol. 30, No. 20, pp.367-370, 2009.

[8] H. Zhang, and J. Wang, "Optimization of sensor array of electronic nose and its application to detection of storage age of wheat grain," *Transactions from the Chinese Society of Agricultural Engineering*, Vol. 22, No. 12, pp.164-167, 2006.

[9] B. Shi, L. Zhao, R. Zhi, and X. Xi, "Optimization of electronic nose sensor array by genetic algorithms in Xihu-Longjing Tea quality analysis," *Mathematical and Computer modelling*, Vol. 58, No. 3-4, pp.752-758, 2013.

[10] M. Tabassum, and K. Mathew, "A genetic algorithm analysis towards optimization solutions," *International Journal of Digital Information and Wireless Communications (IJDIWC)*, Vol. 4, No.1, pp.124-142, 2014.

[11] R. O. Duda, D. G. Stork, and P. E. Hart, "Pattern classification", 2nd ed. New York: Wiley, 2001, pp. 115.

[12] V. V. Sysoev, V. Yu. Musatov, A. V. Silaev, and T. T. Zalyalov, "The optimization of number of sensors in one-chip electronic nose microarrays with the help of 3-layered neural network," *Siberian Conference on Control and Communications SIBCON-2007*, pp. 185-191, 2007.

[13] M. M. Rahman, C. Charoenlarpnopparut and P. Suksompong, "Classification and pattern recognition algorithms applied to E-Nose," *2nd International Conference on Electrical Information and Communication Technology (EICT)*, pp. 44-48, 2015.

[14] M. M. Rahman, C. Charoenlarpnopparut, and P. Suksompong , "Signal processing for multisensor E-nose system: acquisition and classification," *10th International Conference on Information, Communications and Signal Processing*, [CD-ROM], 2015.

[15] Peter Krafft, "Building intelligent probabilistic systems," *https://hips.seas.harvard.edu/blog/2013/02/13/correlation-and-mutual-information*, February 13, 2013. Accessed: April 2, 2016.

Md. Mizanur Rahman received his B.Sc. Engg. (with distinction) in Electronics and Communication Engineering (ECE) from Khulna University (KU), Bangladesh in 2002, M.Sc. in ECE, School of Information Computer and Communication Technology (ICT), from Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU), Thailand in 2014. He is working as a faculty member in ECE, KU, Bangladesh since December 2003. At present he is on study leave from KU, Bangladesh and working as a PhD researcher in ECE, SIIT, TU, Thailand.

Chalie Charoenlarpnopparut received his B.Eng in Electrical Engineering, Chulalongkorn University, Bangkok, Thailand and his M.S and Ph.D in Electrical Engineering from The Pennsylvania State University, PA, USA. He has been an Associate Professor at the School of Information, Computer, and Communication Technology (ICT) at SIIT, Thammasat University since 2002. His research interests are in multidimensional systems and signal processing, robust control, image processing, wavelet and filter bank, signal processing for communication, convolutional code design.

Prapun Suksompong received his B.S. in Electrical and Computer Engineering, M.S. in Electrical and Computer Engineering, and Ph.D. in Electrical and Computer Engineering from Cornell University, USA. He has been an Assistant Professor at the School of Information, Computer, and Communication Technology (ICT) at SIIT, Thammasat University since 2008. His research interest are in wireless communications, indoor positioning principles and localization techniques, computational neuroscience, energy-efficient coding and Poisson process and Poisson convergence.

Pisanu Toochinda received B. Sc. (Chemistry) from Mahidol University, Bangkok, Thailand in 1995. He received his M. Sc. and Ph. D. in Chemical Engineering from The University of Akron, Ohio, USA in 1999 and 2003 respectively. He has been serving as a full-time lecturer at School of Bio-Chemical Engineering and Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University since 2003 where he is currently an Associate Professor in 2016. He was the coordinator of Chemical Engineering program from 2004-2014 and head of the school of Bio-Chemical Engineering and Technology from 2009-2014. His research area are Hydrogen production from alcohol reforming, Photocatalytic synthesis of hydrocarbons from CO_2/H_2O, Gas-solid reactor design, Heterogeneous catalysis, Nano-material /zeolite syntheses, Bio-molecular imprinted materials.