Normalization techniques for gas sensor array as applied to classification for black tea

Bipan Tudu[1], Bikram Kow[2], Nabarun Bhattacharyya[3], Rajib Bandyopadhyay[4]

[1, 2, 4]Department of Instrumentation and Electronics Engineering, Jadavpur University, Salt Lake Campus, Sector III, Block LB, Plot No. 8, Kolkata-700 098, India. Tel.: +91 33 23352587; fax: +91 33 23357254.

[3]Centre for Development of Advanced Computing(C-DAC), E-2/1, Block – GP, Sector – V, Salt Lake, Kolkata-700 091, West Bengal, India. Tel.: +91 33 23576309; fax: +91 33 23575141.

[1]bt@iee.jusl.ac.in, [2]bkiramkow@gmail.com, [3]nabarun.bhattacharya@kolkatacdac.in, [4]rb@iee.jusl.ac.in

Abstract— Assessment of black tea quality is a difficult task due to the presence of a large number of chemical compounds. The present day practice in the tea industry for this purpose is to employ the tea-tasters, who evaluate the quality based on their experience and professional acumen. There is a dire need in the industry to assess the tea quality objectively using instrumental methods. In this pursuit, an electronic nose instrument with five gas sensors has been developed and deployed for declaring tea-taster like scores. It has been observed that pre-processing of gas sensor data improves the classification accuracy and in this paper, a comparative study of different normalization techniques is presented for black tea application using electronic nose. For this study black tea samples were collected from different tea gardens in India. At first Principal Component Analysis (PCA) is used to investigate the presence of clusters in the sensors responses in multidimensional space. Then different normalization techniques were applied on electronic nose data. Finally the comparison of classification accuracy is presented with different normalization techniques using back-propagation multilayer perceptron (BP-MLP) neural network.

Index terms: black tea, electronic nose, gas sensor; normalization technique, principal component analysis, back-propagation multilayer perceptron (BP-MLP).

I. INTRODUCTION

Among all the beverages consumed worldwide, tea is a very popular one and used by all strata of people. But because of the requirement of appropriate climatic conditions, tea plants are available only in a few countries in the world. There are different varieties of tea – green tea, black tea, Oolong tea etc. and out of these, black tea is the most common beverage. For

measurement of tea quality, unfortunately no instrumental methods are deployed on regular basis in the industry and the age-old method of employing professional tea tasters are still being practiced. These tasters, based on their experience and judgment, assess the quality of tea and the pricing of tea is made accordingly. The tea-tasters give a mark in the range of 1 to 10 each for leaf quality, infusion and liquor of the sample [1]. This method is purely subjective and error-prone. Sometimes, the perception of tea quality, even with the same sample, varies from one taster to another and thus creates a lot of confusion. Thus there is a demand in the industry to have low-cost, portable solutions for quality evaluation of black tea. In this regard, electronic nose has been demonstrated to be an appropriate candidate [2] for the same.

An electronic nose is an instrument, which comprises of a sensor array, a suitable odour delivery system, data acquisition and processing module and a pattern classifier. The block diagram is shown in figure 1. When a sample containing volatile organic compounds is presented to the sensor array through the odour handling mechanism, a signature is generated by the sensor array which is characteristic of its odour and the pattern classifier deciphers the signature. The pattern classification algorithms and data processing techniques are critical components in an electronic nose and are responsible for the implementation and successful commercialization of electronic nose systems. For a particular application, the pattern classifier is first trained with a database of known samples and then this trained classifier is used to predict the response for an unknown sample.

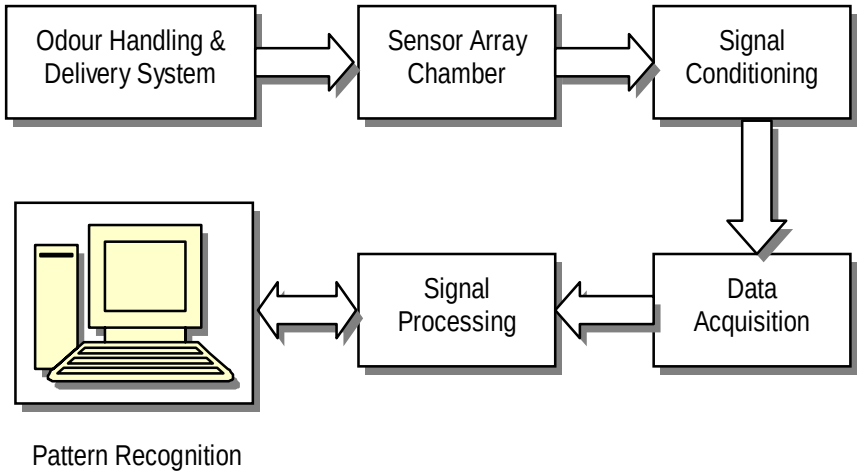

Figure 1. Block diagram of electronic nose

Good performance of pattern classifier is an essential requirement for a useful electronic nose instrument. Associated with the classifier, the normalization techniques are equally important and proper choice of the normalization technique may enhance the performance of the pattern recognition algorithm and the overall electronic nose system as well. Some of the interesting studies on data preprocessing and normalization have been presented in [3], [4]. In our study, an electronic nose set-up has been developed with five tin-oxide sensors from FIGARO, Japan for black tea quality evaluation. A variety of tea samples are collected from different tea gardens in India and each of these samples is subjected to the electronic nose instrument and an experienced tea-taster as well. Then several standard normalization techniques have been applied on the electronic nose data. The back-propagation multilayer perceptron (BP-MLP) algorithm is used with the normalized data and the comparison of classification accuracy is presented for all the normalized techniques.

II. EXPERIMENTAL SET-UP

a. Customized Electronic Nose Set-up for Black Tea

A customized electronic nose set-up for quality evaluation of black tea has been developed and is shown in figure 2. Specially designed sample holders made of glass have been used for the experimental runs. The glass sample holders may be fixed to the instrument by simple threaded fitting. For black tea, an array of Metal Oxide Semiconductor (MOS) sensors has been used for assessment of volatiles in the set-up. From the response sensitivity of individual sensors, a set of five gas sensors from Figaro, Japan (TGS-832, TGS-823, TGS-2600, TGS-2610 and TGS-2611) has been selected for odour capture from black tea [1]. The MOS sensors are conductometric in nature, and their resistance decreases when subjected to the odour vapour molecules. The change in resistance with respect to their original values is converted into voltage and then taken to the PC through analog to digital converter cards for subsequent analysis in the computational model. The outputs of the sensors are acquired in the PC through the data acquisition card USB 6008 from National Instruments®.

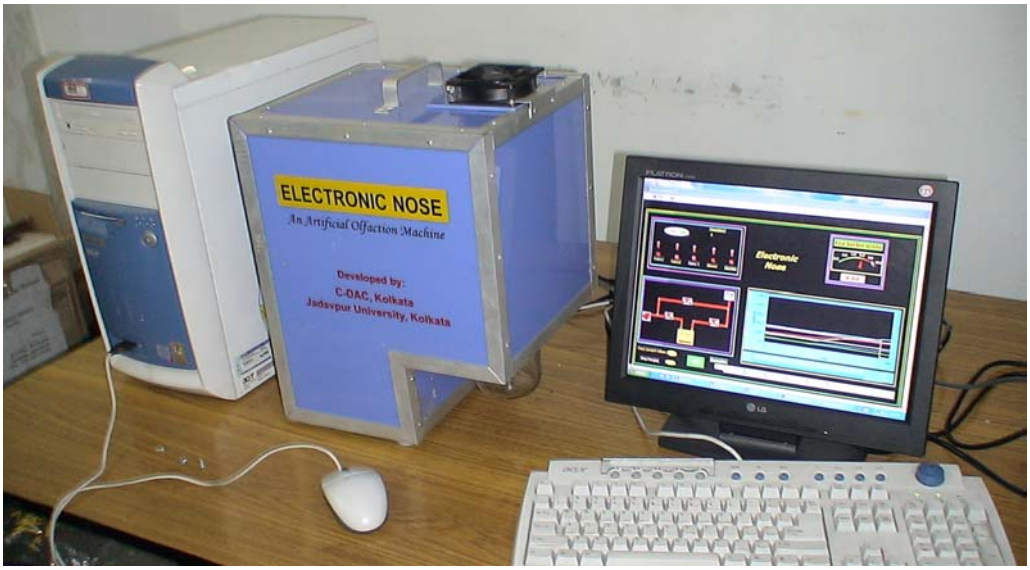

Figure 2. Customized electronic nose set-up developed

The experimental sniffing cycle consists of automated sequence of internal operations: (i) headspace generation, (ii) sampling, (iii) purging and (iv) dormancy before the start of the next sniffing cycle. Headspace generation ensures adequate concentration of volatiles released by tea within the sample holder by blowing regulated flow of air on the sample. During sampling, the sensor array is exposed to a constant flow of volatiles through pipelines inside the electronic nose system. Data from all the sensors are stored all through this sampling operation, but the steady-state value for each sensor is considered for the computation. During purging operation, sensor heads are cleared with blow of fresh air so that the sensors can go back to their baseline values. The programmable time dormancy cycle is the suspended mode of the electronic nose between two consecutive sniffing cycles.

The PC-based data acquisition and automated operation of all these cycles are controlled by specially designed software. The software has been developed in LabVIEW® of National Instruments. The software design is focused on features like minimum operator intervention in the production floor and user-friendly graphical user interface, so that operators with basic computer literacy can handle the instrument.

The experimental conditions for black tea classification are given as follows:

Amount of black tea sample = 50 grams

Temperature = 60° ±2°C

Headspace generation time = 30s

Data collection time = 100s

Purging time = 100s

Airflow rate = 5 ml/s

The above experimental conditions have been optimized for black tea quality evaluation on the basis of repeated trials and sustained experimentation.

b. Sample Collection

Samples are collected from four tea gardens in India spread across eastern and north-eastern tea heartland throughout the tea production season (i.e. from March to December). The respective gardens have provided one batch of sample per week. In total, 194 samples have been collected for this study and all these samples have been subjected to evaluation by a tea taster. Table 1 shows a sample of tasters' mark evaluation sheet for black tea samples from different gardens.

Table 1: Taster evaluation sheet

Sample code	**Average scores (1 to 10)**		
	Leaf quality	**Infusion**	**Liquor**
KON050604-01	7	5	3
KON020904-01	6	5	5
KON071204-01	5	4	4
KON210904-01	7	5	6
MAT070504-01	8	8	8
FUL150604-01	7	6	5
GLN180604-01	8	8	7
MAT100604-01	7	6	7
GLN180604-02	8	7	7

While leaf quality and infusion scores are based on visual inspection of the samples by the tasters, the marks given against "liquor" is the combined perception of aroma and flavour of the sample. The scores assigned to liquor, therefore, have been considered by us for training the neural network model.

III. DATA ANALYSIS

In the present electronic nose based study on black tea, each sniffing cycle produces a huge matrix with five columns corresponding to five sensors in the array. Each row contains the response at various sampling instants. In the system developed by us, each sniffing cycle consists of headspace generation (30 seconds), sampling (100 seconds) and purging (100 seconds) operations. The sensor data was acquired by the computer during headspace generation and sampling cycles only in an automated sequence of operation. Clearly, the sensor outputs are at their baseline during the headspace generation and significant variation in sensor outputs are observed during the sampling cycle. Therefore, the data matrix stored in the computer in each sniffing cycle will consist of a mixture of both baseline as well as actual sensor responses as shown in figure 3.

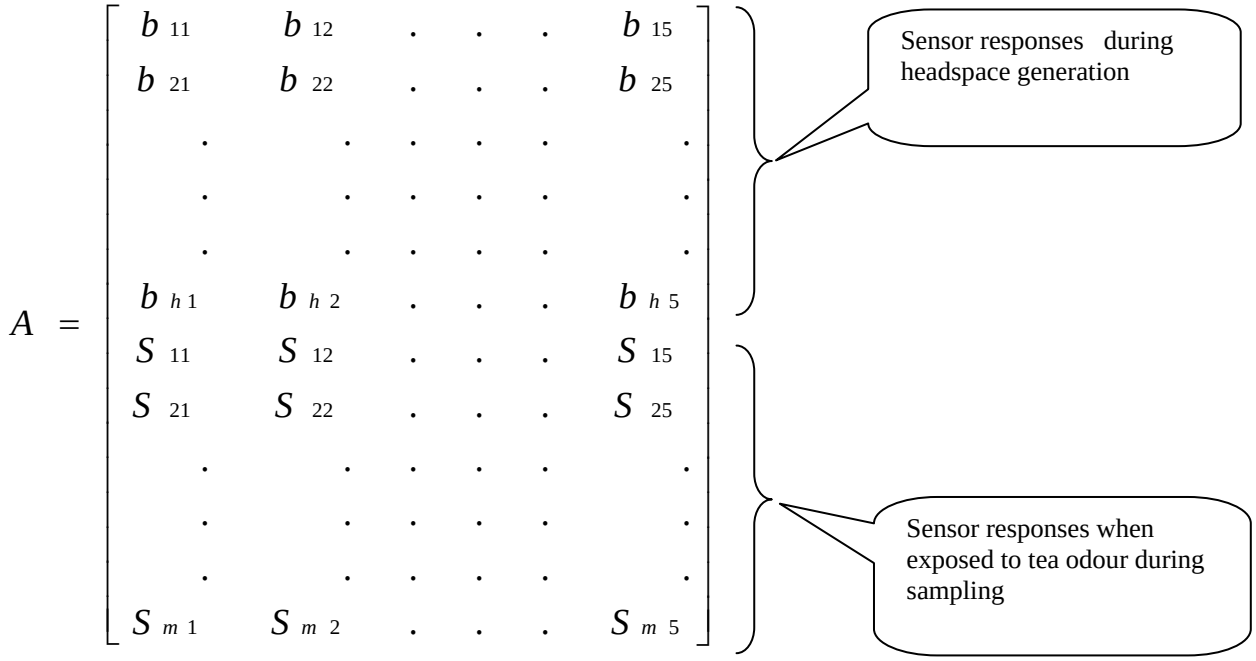

Figure 3. Data matrix formed out of sensor array output

In the above data matrix, the segment b_{11} to b_{h5} represents baseline responses of the sensors during headspace generation and S_{11} to S_{m5} represents the sensor responses when exposed to tea odour during sampling. In our study, maximum value of S_{ij} for each column has been found out and a vector M is formed with these maximum values.

So, $M = [S_{i1\max} \ldots\ldots\ldots\ldots S_{i8\max}]$.

For each run, the maximum value vectors associated with each sniffing cycle are subtracted from base values and a separate matrix *N* is formed. While the number of columns of this matrix would be equal to the number of sensors in the sensor array, (which is 5 for our case) the number of rows will depend on the sample runs.

a. Normalization Techniques

Data collection is the initial step for data investigation. The gas sensors generate electrical signals when subjected to the odour molecules and the signals are converted to numeric data using analog-to-digital converters. These data are then used for computation and finally for classification. This procedure often causes the difficulty in the classification problem as the original signal may be distorted due to the characteristics and limitations of the transducer. In the context of electronic nose, the signals, usually considered, are the maximum of the sensors responses, and these are the raw or non-normalized data. Prior to classification using a suitable pattern recognition algorithm, these raw data are normalized [5]. An appropriate normalization technique may improve the pattern classification system in an electronic nose, but there are no general guidelines to determine the appropriate normalization technique. Table 2 shows several standard normalization techniques [6-7].

Table 2: Standard data normalization techniques

Normalization	**Mathematical expression**
Range scale$_1$	$A_{ij} = \left(A_{ij} - \min(A_j)\right) \Big/ \left(\max(A_j) - \min(A_j)\right)$
Range scale$_2$	$A_{ij} = \left(2(\frac{(A_{ij} - \min(A_j))}{(\max(A_j) - \min(A_j))}) \right) - 1$
Relative scale$_1$	$A_{ij} = A_{ij} \Big/ \max(A)$
Relative scale$_2$	$A_{ij} = A_{ij} \Big/ \max(A_j)$
Baseline subtraction	$A_{ij} = A_{ij} - A_{1j}$
Global method (sensor auto scaling)	$A_{ij} = (A_{ij} - mean(A)) \Big/ std(A)$
Local method	$A_{ij} = A_{ij} \Big/ \sqrt{\sum_1^n A_{ij}}$

In Table2, seven normalization techniques are presented, and terms used are explained below.

A is the feature matrix for n samples from p sensors,

A_{ij} is the i^{th} sample of the j^{th} sensor,

A_j contains all n responses samples for sensor j,

and A_i contains all p responses for the sensors at the i^{th} sample.

Different normalization techniques bear different meanings [8], as like Relative scale_1 gives a global compression of values with a maximum value of 1, Relative scale_2 compresses values per feature with a maximum value of 1. Auto scale sets the mean at the origin and the variance within the data to 1, often used when responses are on different magnitude scales. Range scale_1 and Range scale_2 set the limits at [0, 1] and [-1, 1] respectively. A baseline subtraction removes the base reading of a sensor, often used in temporal data collection.

IV. RESULTS AND DISCUSSION

Experimentations with electronic nose have been performed with 194 finished tea samples with six different taster scores for the samples and the corresponding sensor output signatures in the computer. Initially, PCA analysis was carried out on raw data set as well as on normalized date set. Further, BP-MLP was tested on normalized data set. Different normalization techniques and BP-MLP algorithm were carried out on MATLAB® platform.

a. Data Processing using PCA

The multi-dimensional data matrix formed out of the signatures from the multi-sensor array has been subjected to Principal Component Analysis (PCA) [9]. PCA plots obtained before preprocessing and after preprocessing with each of the normalization techniques are shown figures 4(a) to 4(h). It can be observed from the PCA plots that the formations of clusters for samples belonging to a particular taster mark are more distinct with normalized than data without normalization.

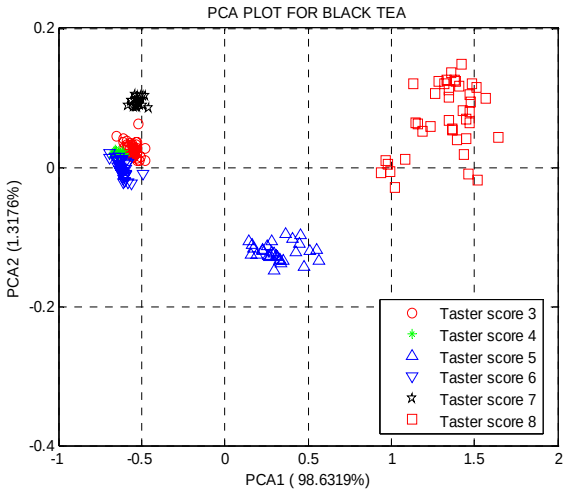

Figure 4(a). PCA plot before normalization

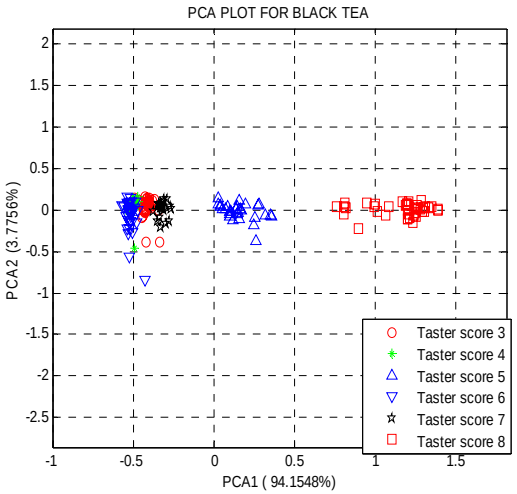

Figure 4(b). PCA plot after normalization using range scale$_1$

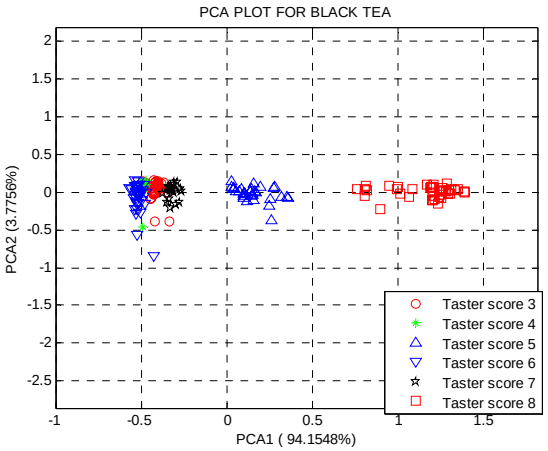

Figure 4(c). PCA plot after normalization using range scale$_2$

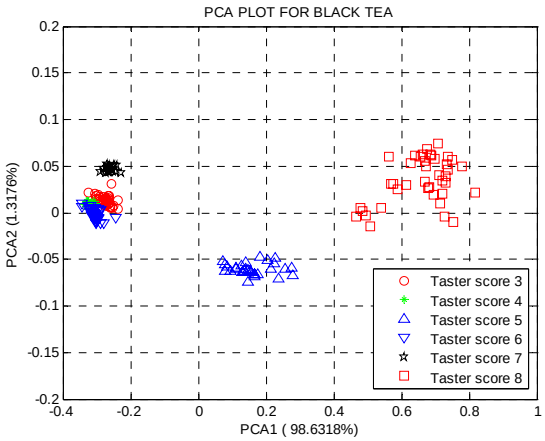

Figure 4(d). PCA plot after normalization using relative scale$_1$

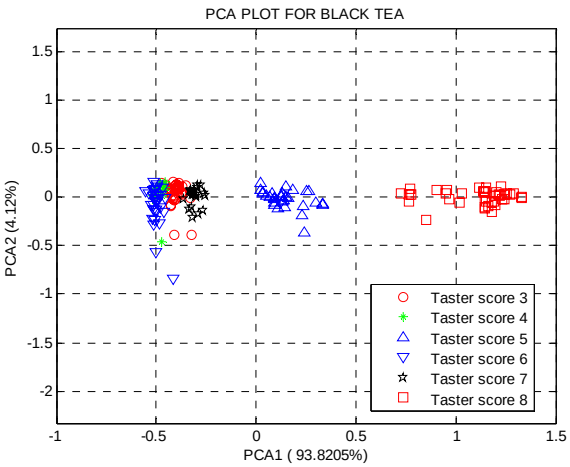

Figure 4(e). PCA plot after normalization using Relative scale$_2$

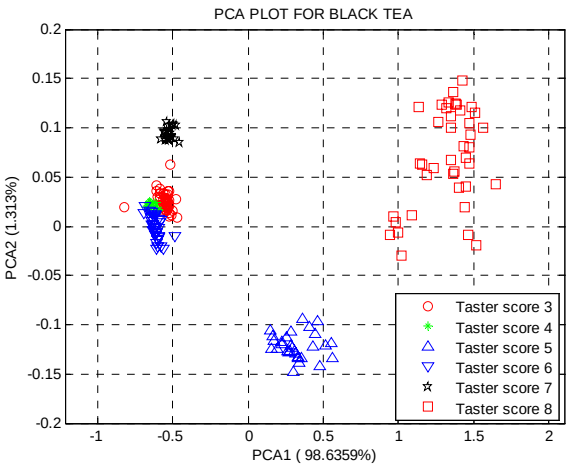

Figure 4(f). PCA plot after normalization using baseline subtraction

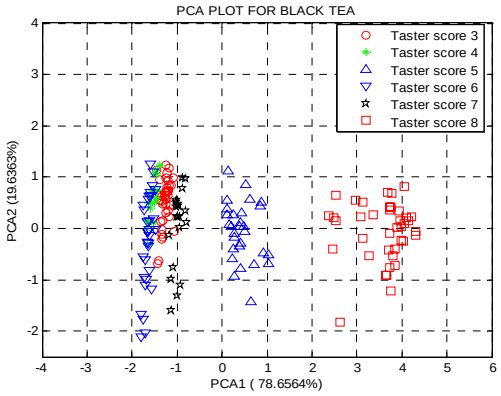

Figure 4(g). PCA plot after normalization using sensor auto scaling

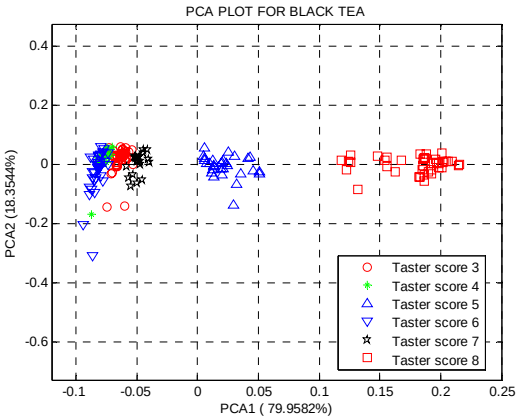

Figure 4(h). PCA plot after normalization using local method

b. Classification of Normalized Data using BP-MLP

A three-layer Back propagation Multilayer Perceptron (BP-MLP) model with one input layer, one hidden layer and one output layer has been considered [10]. The input layer has been fed with the output from the sensor array processed data and the output layer has been configured to show the taster's score as shown in figure 5. For five sensors in the multi-sensor array, the ANN has five input nodes. It has been observed that the taster's score vary within a range from 3 to 8. So the output layer of the network needs to be assigned with six nodes as shown in figure 5. Convergence during learning process has been obtained with acceptable accuracy

with only one hidden layer with 8 nodes. Comparative results with the normalization techniques described above are presented in Table 3.

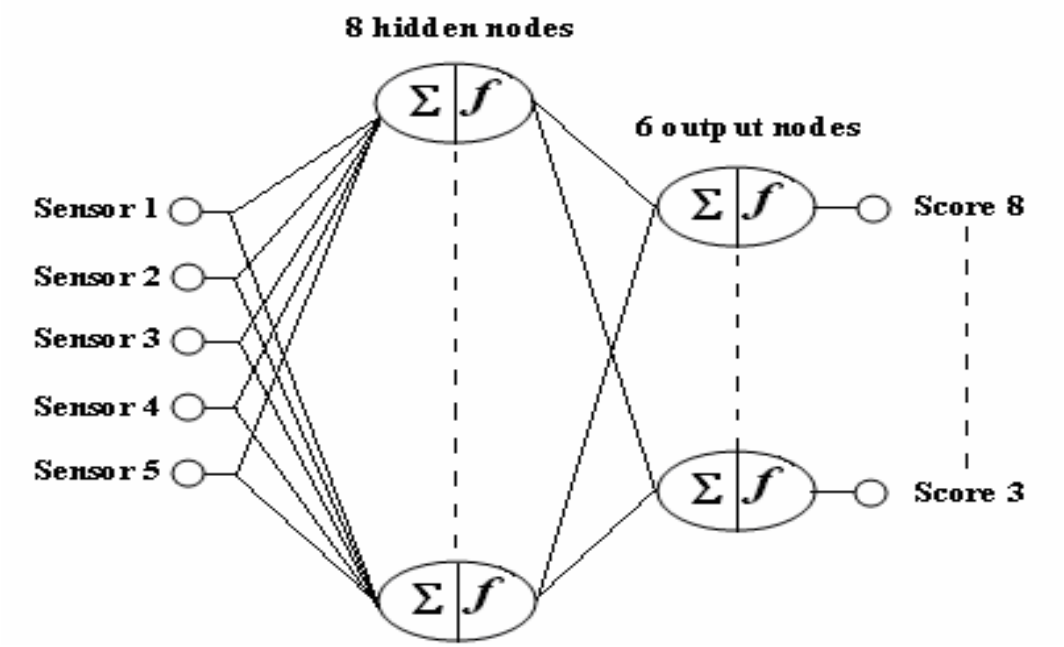

Figure 5. Three-layer BP-MLP network

When raw data set was tested using BP-MLP algorithm, the classification accuracy achieved was 60.25%. But on normalized data set, the classification accuracy improves. Results obtained for different normalization techniques using BP-MLP are shown in Table 3. It is to be noted from this study that for classification of black tea using electronic nose based on BP-MLP algorithm, normalization technique of Range $scale_2$ gives the best classification rate, which is more than 93%.

Table 3: Results obtained for different normalization techniques

Normalization techniques	**BP-MLP architecture**	**Correct classification with training data (%)**
Range $scale_2$	Input node =5	93.814
Range $scale_1$	Hidden node = 8	91.237
Relative $scale_2$	Output node = 6	90.206
Baseline subtraction	Epoch =1400	83.505
Relative $scale_1$	Learning rate for hidden layer = 0.5 learning rate for output layer = 0.1	80.414
Global method		90.722
Local method		60.825

V. CONCLUSION

In this study, our objective was to carry out a comparative study between different normalization techniques as applied for black tea aroma classification using electronic nose. In total, eight normalization methods have been considered. The performance of each of the normalization techniques are assessed by the evaluation of the classification accuracy on electronic nose data for 194 black tea samples. The pattern classification algorithm used is BP-MLP. It is observed that the normalization techniques influence the results considerably and careful choice of the appropriate method is very important. Moreover, as there are no general guidelines for the selection of the most suitable method of normalization, this can be done with the practical data only. All in all, it may be concluded that for an electronic nose to be a successful instrument for a particular application, choice of normalization technique is important and should be done with application-specific field data prior to its deployment in practice.

REFERENCES

[1] N. Bhattacharyya, R. Bandyopadhyay, M. Bhuyan, B. Tudu, D. Ghosh, and A. Jana, "Electronic nose for black tea classification and correlation of measurements with

"Tea Taster" marks", IEEE Trans. Instrum. Meas., Vol. 57, No. 7, pp. 1313-1321, Jul. 2008.

[2] R. Dutta, E. L. Hines, J. W. Gardner, K. R. Kashwan, and M. Bhuyan, "Tea quality prediction using a tin oxide-based electronic nose: An artificial intelligence approach", Sens. Actuators B, Vol. 94, pp. 228-237, Sep. 2003.

[3] M. Padro, G. Niederjaufner, G. Benussi, G. Faglia, G. Sberveglieri, M. Holmberg, and I. Lundstrom, "Data preprocessing enhances the classification of different brands of Espresso coffee with an electronic nose", Sens. Actuators B,Chem., Vol. 69, pp 397-403, 2007.

[4] R. G. Osuna and H. T. Nagle "A method for evaluating data –preprocessing techniques for odor classification with an array of gas sensors", IEEE Trans. Syst. Man Cybernet., Part B, Vol.29, pp. 626-632, 1999.

[5] Bipan Tudu, Nabarun Bhattacharyya, Bikram Kow, and Rajib Bandyopadhyay, "Comparison of Multivariate Normalization Techniques as Applied to Electronic Nose Based Pattern Classification for Black Tea", in Proc. IEEE 3rd International Conference on Sensing Technology (ICST 2008), Tainan, Taiwan, Nov.30- Dec.3 , 2008, pp. 254-258.

[6] P. C. Jurs, G. A. Bakken, and H. E. McClelland, "Computational methods for the analysis of chemical sensor array data from volatile analytes", Chemical Rev., Vol. 100, pp. 2649- 2678, 2000.

[7] J. W. Gardner, P. N. Barlett, *Electronic noses: Principles and Applications,* Oxford University Press, 1999.

[8] S. M. Scott, D. James, and Z. Ali, "Data analysis for electronic nose systems", Microchim Acta, Vol. 156, pp. 183–207, 2007.

[9] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, *Singular value decomposition and principal component analysis, in A Practical Approach to Microarray Data Analysis,* D. P. Berrar, W. Dubitzky, and M. Granzow (Eds.), Norwell, MA: Kluwer, pp. 91-109 ch. 5, 2003.

[10] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed. Hong Kong: Pearson Educ. Asia, 2001.