ELSEVIER

Identification of *Staphylococcus aureus* infections in hospital environment: electronic nose based approach

Ritaban Dutta [a, *], David Morgan [b], Nicky Baker [b], Julian W. Gardner [a], Evor L. Hines [a]

[a] *School of Engineering, University of Warwick, Coventry CV47AL, UK*
[b] *Birmingham Heartlands Hospital, Bordesley Green East, Birmingham B95SS, UK*

Abstract

An electronic nose (e-nose), the Cyrano Sciences' Cyranose 320 (C-320), comprising an array of 32 polymer carbon black composite sensors has been used to identify two species of *Staphylococcus aureus* bacteria, namely methicillin-resistant *S. aureus* (MRSA) and methicillin-susceptible *S. aureus* (MSSA) responsible for ear nose and throat (ENT) infections when present in standard agar solution in the hospital environment. C-320 e-nose has also been used to identify coagulase-negative staphylococci (C-NS) in the hospital environment. Swab samples were collected from the infected areas of the ENT patients' ear, nose and throat regions. Gathered data were a very complex mixture of different chemical compounds. An innovative object-oriented data clustering approach was investigated for these groups of *S. aureus* data by combining the principal component analysis (PCA) based three-dimensional scatter plot, Fuzzy C Means (FCM) and self-organizing map (SOM) network. Using these three data clustering algorithms simultaneously better 'classification' of three bacteria subclasses were represented. Then three supervised classifiers, namely multi-layer perceptron (MLP), probabilistic neural network (PNN) and radial basis function network (RBF), were used to classify the three classes. A comparative evaluation of the classifiers was conducted for this application. The best results suggest that we are able to identify three bacteria subclasses with up to 99.69% accuracy with the application of the RBF network along with C-320. This type of bacteria data analysis and feature extraction is very difficult. But we can conclude that this preliminary study proves that e-nose based approach can provide very strong solution for identifying *S. aureus* infections in hospital environment and early detection.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Electronic nose (e-nose); Cyrano Sciences' Cyranose 320 (C-320); *Staphylococcus aureus*; Methicillin-resistant *S. aureus* (MRSA); Methicillin-susceptible *S. aureus* (MSSA); Coagulase-negative staphylococci (C-NS); Principal component analysis (PCA); Fuzzy C Means (FCM); Self-organizing map (SOM); Multi-layer perceptron (MLP); Probabilistic neural network (PNN); Radial basis function network (RBF)

1. Bacterial subclasses for this study

Methicillin-resistant *Staphylococcus aureus* (MRSA) are a subgroup within a group of organisms known as *S. aureus*. MRSA are characterized by their resistance to treatment with commonly used antibiotics, in contrast to the remainder of the *Staphyloccocus aureus* group which are referred to as methicillin-susceptible *S. aureus* (MSSA). Both MRSA and MSSA can cause infection but individuals may also carry the organism without being infected by it. An individual, who carries the organism, but is not infected, is said to be a 'carrier' or 'colonized'. At any one time up to 33% of healthy individuals carry *Staphyloccocus aureus*, including MRSA, predominantly in their noses and also at other sites (Fig. 1).

Staphyloccocus aureus can give rise to infections varying from mild, e.g. boils and infected cuts, to severe, e.g. infections of bones, lungs, heart and blood stream. The difficulty in treating MRSA with antibiotics has led to concern about this particular group of staphylococcal infections. All strains of *Staphyloccocus aureus*, including MSSA and MRSA, are capable of causing hospital-acquired infection.

Organisms can be passed to patients from contact with hands or directly from the environment. The latter includes

* Corresponding author. Tel.: +44 2476 575952; fax: +44 2476 418922.
 E-mail addresses: dwmorganfrcs@btopenworld.com (D. Morgan),
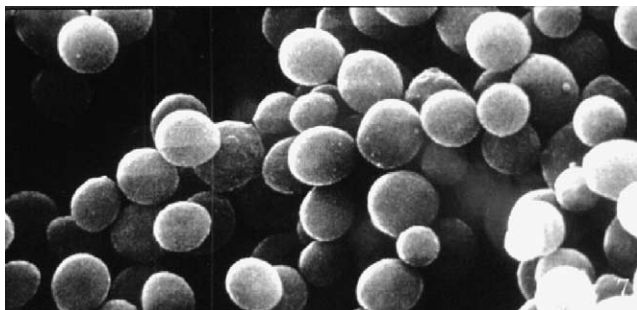r.dutta@warwick.ac.uk, ritaban.dutta@warwick.ac.uk (R. Dutta).

Fig. 1. Scanning electron micrograph of *S. aureus*.

air, dust, clothing, soft furnishings, surfaces and equipment. Some patients are more susceptible to colonization and infection than others. These include the elderly, patients with wounds, ulcers or bedsores, catheterized patients, those who have received antibiotics and those who have been, or who are, hospitalized or institutionalized.

Coagulase-negative staphylococci (C-NS) is among the most commonly isolated bacteria in clinical microbiology laboratories. Such coagulase-negative, novobiocin-susceptible staphylococci as *Staphylococcus epidermidis* have emerged as a major cause of infection, particularly in hospitalized patients with indwelling foreign bodies and in immunocompromised patients.

Coagulase-negative staphylococci (C-NS) are currently both the most common cause of true nosocomial bacteremia and the most common isolate of falsely positive blood cultures. Of the 31 recognized species of C-NS, two are responsible for the majority of C-NS infections: *S. epidermidis*—the most common cause of both foreign device infection and nosocomial bacteremia; and *S. saprophyticus*—second only to *Escherichia coli* as a cause of acute urinary tract infections (UTI) in young, healthy, sexually active women. *S. epidermidis* bacteremia, with its 10–34% mortality rate, is usually preceded by skin colonization with and displacement of normal skin flora by *S. epidermidis*. The symptoms of UTI due to *S. saprophyticus* are indistinguishable from those of *E. coli*, and both pathogens are susceptible to most antibiotics used to treat UTIs. However, with the exception of *S. saprophyticus*, >80% of C-NS infections are resistant to b-lactam antibiotics, erythromycin, clindamycin, tetracycline, chloramphenicol, trimethoprim/sulfamethoxazole, and aminoglycosides [1].

Although the pathogenesis of C-NS infections is controversial, there are three major factors which appear important: skin and sebaceous or apocrine gland colonization; bacterial adherence to foreign devices; and production of slime (an exopolysaccharide which enhances adherence to surfaces and has anti-phagocytic properties). Approximately 75% of patients with C-NS bacteremia have an indwelling medical device (usually an intravascular catheter). Criteria for distinguishing true bacteremia from contamination includes semi-quantitative culture of catheter tips (>15 colony forming units (CFU) indicate true infection), or demonstrating a 5–10-fold

higher number of CFUs on quantitative blood cultures drawn through the catheter lumen as compared to blood cultures obtained via a peripheral vein. The properties which allow C-NS to infect medical devices result in C-NS being the most common cause of prosthetic valve endocarditis (C-NS only cause 1–3% of native valve endocarditis) and nosocomial bacteremia (at 30,000 cases/year), as well as a major cause of vascular graft, central nervous system shunt, prosthetic joint, peritoneal dialysis catheter, and postsurgical endophthalmitis infections. At present the effective therapy for methicillin-resistant C-NS is vancomycin + rifampin + gentamicin.

2. Electronic nose system

The electronic nose (e-nose) is an instrument that has been developed as a simplified model of the human olfactory system. Gardner and Bartlett defined an electronic nose as "An instrument, which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern recognition system, capable of recognizing simple or complex odours".

The EN system is designed for automated detection and classification of odours, vapours, and gases. It can also perform simple odour discrimination and provide measurement of odour intensity. The two main components of a C-320 e-nose are the sensing system and the automated pattern recognition system. In this paper we describe the use of a commercial handheld electronic nose, the Cyrano Sciences' C-320 as an effective tool for illness diagnosis (see Fig. 2). The instrument settings, defined methods and raw data can

Fig. 2. Typical Cyranose 320 e-nose which contains 32 carbon black conducting polymer composite sensors.

be swapped, stored and further processed on Windows-based PC using PC nose software.

The system consists of 32 individual carbon black polymer composite resistive sensors configured into an array. When the sensors are exposed to vapours or aromas they swell, reducing conductance between the carbon sensors. This in turn increases the total resistance of the film, which is monitored as the sensor signal [2]. The responses from all sensors in the array form a response pattern or "smell-print". The sensor technology yields a distinct response signature for each vapour regardless of its complexity and produces.

3. Swab samples

Swab samples were collected from the ENT patients who were suffering from these bacteria infections. The *S. aureus* infected swab sample experiments were conducted using C-320 (carbon black conducting polymer composite sensors based e-nose, a commercial e-nose) at the Heartland hospital, Birmingham, UK. The Heartland hospital, Birmingham, UK ethic committee approved this study.

With the assistance of two ENT specialists in Heartlands hospital swab samples from three types of patients' swab samples were collected. These three types of samples were as follows:

- *Sample* 1: Swab samples representing methicillin-resistant *S. aureus* (MRSA).
- *Sample* 2: Swab samples representing methicillin-susceptible *S. aureus* (MSSA).
- *Sample* 3: Swab samples representing coagulase-negative staphylococci (C-NS).

We got 50 different swabs from 50 different patients for each type of bacterial subclass and all samples were sniffed following same procedure. Effectively 150 patients swab samples were studied using C-320.

4. Test procedure

A strategy to collect the ENT odour samples and sampling procedure was agreed with the ENT specialists and was used throughout the data gathering process. Both protocol and sampling technology were kept simple to be cost effective and non-application specific [4,5]. ENT patients were collected for subsequent analysis using the Cyranose. Swab samples were collected from the infected areas of the ENT patients' ear, nose and throat regions from the clinical patients. After collection, all swab samples were kept in ISO S type agar solution in typical 15 cm^3 vials. This vial size was chosen so as to generate sufficient headspace from the bacterial solution for sniffing purposes. All swab samples were kept in the agar solution in a vial to help the growth of the ENT bacteria present in the ENT patients' swab samples.

Later we used C-320 electronic nose system to sniff all the samples. We put approximately 3 mg of agar solution of bacteria in a 10 ml vial. We kept the vial containing bacterial solution for 5 min; it was to generate some headspace from bacterial solution. C-320's inlet was inserted into the vial and headspace was sniffed.

The general method for data collection protocol was as follows:

1. Collect two separate swab samples from the infected area of patients.
2. Collect two samples from each patient; one for standard bacteriological tests and another one were kept in a culture medium for 24 h, and the last one was used for non-cultured direct sniffing using the Cyranose e-nose.
3. Keep the non-cultured sample in a closed vial (15 ml) for 5 min.
   This time will allow the bacteria present to produce sufficient volatile aromatic compound in the vial for odour sniffing.
4. Sniff the same vial directly by using Cyranose 320 e-nose 10 times.
5. Repeat and store 10 'sniffing' from each patient's swab sample using the Cyranose 320 e-nose (this repetition for collecting data from the same patient will allow us to build a large data base for our subsequent data analysis) [6].

5. Data visualization

To visualize EN data, graphical representation is plotted traditionally.

Surface plot representation could be helpful way to represent the data. Using the 'surfc' function of MATLAB 6.1 [7], all the surface plots of all data sets have been generated. 'Surfc' function is used to view mathematical functions over a rectangular region. 'Surfc' creates coloured parametric surfaces specified by three different components, such as, normalized values of sensors' responses, number of sensors and number of gathered data points.

If the normalized data matrix is 'Z', then surfc(Z) creates a three-dimensional shaded surface from the z components in matrix Z, using x = 1:n and y = 1:m, where [m, n] = size(Z). The height, Z, is a single-valued function defined over a geometrically rectangular grid. Z specifies the colour data as well as surface height, so colour is proportional to surface height. See Fig. 3.

6. Signal pre-processing

The choice of the data pre-processing algorithm has been shown to affect the performance of the pattern recognition stage. All data was normalized using a fractional difference model: $dR = (x_{ij} - x_i^{min})/(x_i^{max} - x_i^{min})$ where R is the response of the system to the sample gas, the reference gas
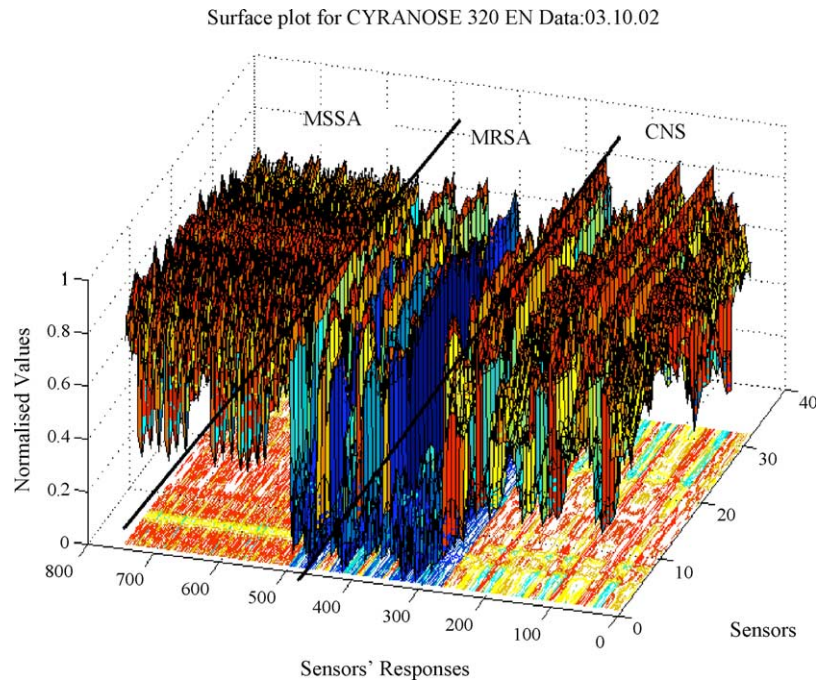
Fig. 3. Representative surface plot for ENT bacteria samples' e-nose data gathered using a Cyranose 320 e-nose system. All sensors' responses are normalized to be the range of [0, 1].

being the ambient room air. The complete bacteria data set was then normalized, by dividing each dR by the maximum value for each sensor, in order to set the range of each sensor to [0, 1].

7. Conventional exploratory technique for data clustering

The use of principal component analysis (PCA) [8], Fuzzy C Means (FCM) [9] and self-organizing map (SOM) [10] to assess clustering within the data set is now discussed. These exploratory techniques are used to investigate how the data cluster in the multi-sensor space. Several techniques were applied to verify that the categories established by each were not arbitrary and the groups formed match the three types of ENT bacteria classes. The objective of this analysis was to establish simple classes for the different bacteria species in order to examine whether or not the data clusters could be separated in preparation for the pattern recognition stage.

8. Results

8.1. Results from PCA

PCA is an effective linear method for discriminating between the e-nose responses to simple and complex odours. The method consists of expressing the response vectors in terms of a linear combination of orthogonal vectors that account for a certain amount of variance in the data. Fig. 4 shows

the results of applying PCA to an array of 32 polymer sensors when applied to the three different types of swab samples of *S. aureus* infections in hospital environment. Since polymer sensors respond in a similar manner, over 99.59% of the variance is described by the first two PCs. Three different clusters or categories are apparent and are associated to three types of samples of three different bacteria species. Three PCs were kept, which accounted for 99.86% of the variance in the dataset (PC #1, PC #2 and PC #3 accounted for 99.12, 0.49 and 0.15% of the variance, respectively). It is also clearly evident that the sensors are highly linearly correlated. Most of the variance in the data is explained by considering only the first principal component (PC1), which implies that the sensor responses are highly correlated. As PC1 accounts for most of the information in the data, this suggests that the clusters were not made any more evident using PCA. That is linear PCA analysis is not informative for this type of data. The objective of this analysis was to establish simple classes for the different bacteria species in order to examine whether or not the data clusters could be separated, prior to the conventional pattern recognition stage. Manually we have drawn three approximate cluster areas using line around individual group data points. These lines were drawn just to indicate probable cluster position more clearly.

8.2. Results from FCM analysis

Fuzzy C Means (FCM) is an unsupervised method. Clustering essentially deals with the task of splitting a set of patterns into a number of more-or-less homogeneous classes (clusters) with respect to a suitable similarity measure such
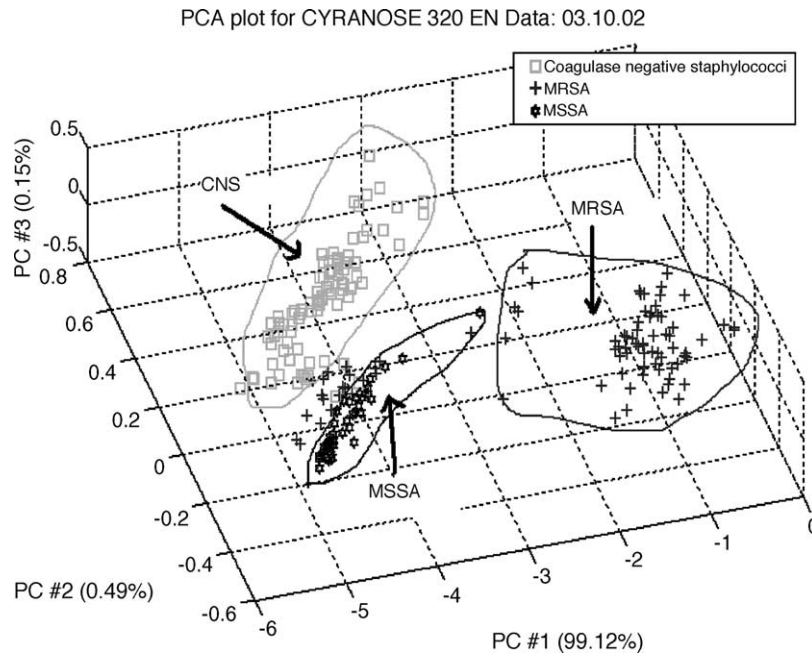
Fig. 4. Three-dimensional view of the PCA plot for the e-nose data sets from three different ENT bacteria swab samples. In this analysis the first PC describes over 99.12% of the variance in the data where as second PC describes 0.49%.

that the patterns belonging to any one of the clusters are similar and the patterns of different clusters are as dissimilar as possible.

The similarity measure used has an important effect on the clustering results since it indicates which mathematical properties of the data set should be used and in what way they should be used in order to identify the clusters. FCM clustering provides partitioning results with additional information supplied by the cluster membership values indicating the degrees of belongingness, where *C* is the total number of

clusters. In Fig. 4 possible cluster centres have been depicted and marked. Three types of ENT bacteria data have been estimated using FCM technique and they are marked inside the original data points in Fig. 5.

8.3. Results from SOM analysis

The self-organizing map (SOM) algorithm was developed by Kohonen to transform an incoming signal pattern of arbitrary dimension into a one or two-dimensional discrete map.
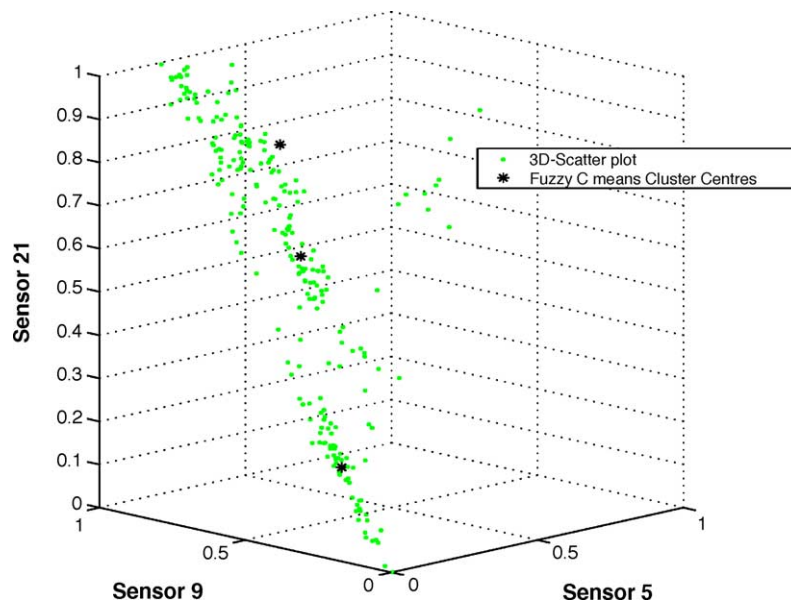

Fig. 5. Three-dimensional view of the unsupervised FCM analysis for three types of ENT bacteria data from non-conducting polymer sensor based Cyrano's EN.

Analyses carried out using Kohonen ANNs fall into the category of "unsupervised learning", in which the relevant multivariate algorithms seek "clusters" in the data; unsupervised learning allows the researchers to group objects together on the basis of their perceived closeness in *n*-dimensional hyperspace (where *n* is the number of variables or observations made on each object). SOM in some sense are quantitative, but are better seen as qualitative since its main purpose is to distinguish objects or classify different classes among the populations. SOM is more closely related to the neural structure of the human olfactory cortex than other ANN because it emulates parts of the brain. SOM applied to e-nose system typically utilizes a single two-dimensional layer of neurons in addition to an input layer of branched nodes. If the system is left alone in an environment of interest, the network learning algorithm processes the sensor outputs step by step, and constructs an internal representation of the environment.

SOM accumulate a lot of statistical information in an unsupervised fashion and benefit to e-nose systems because of their inherent ability such as dimensionality reduction.

The SOM method was applied to the ENT data set in order to investigate clustering using the responses from the 32 polymer sensors. A SOM network of [3 × 1] dimensions was created and trained with the entire ENT data set. Subsequently samples are associated with one of the output neurons and neurons are grouped together to form categories of each identified class of ENT. In the bottom of Fig. 6 there are three neurons which indicate the initial weights of the SOM neurons before training. The trained SOM neurons have got their probable positions inside the data according to the positions

of the three clusters or three types of ENT bacteria groups after 500 training cycles.

So after training of the SOM network the trained SOM neurons locate their probable positions inside the data according to the positions of the most probable clusters; where the three-dimensional weight coordinates of the SOM neurons get new values which are completely different from the initial weights of the SOM neurons before training.

9. Combined SOM, FCM and PCA analysis: a new approach

SOM and FCM were applied to the data set in order to investigate clustering using the responses from the 32 sensors. A SOM network is a non-linear artificial neural network (ANN) paradigm, which is able to accumulate statistical information about data with no other supplementary information than that provided by the sensors. Various SOM networks were created and trained with the entire data set, subsequently samples were associated with one of the neurones and neurones were grouped together to form categories corresponding to each identified bacteria [3].

An innovative data clustering approach was investigated for these bacteria data by combining the three-dimensional PCA scatter plot, FCM and SOM network. This is depicted in Fig. 7. In multisensor space, normalized data sets were represented using three-dimensional scatter plots. From the FCM approach, a cluster centre is found for each group by minimising a dissimilarity function. These cluster centres
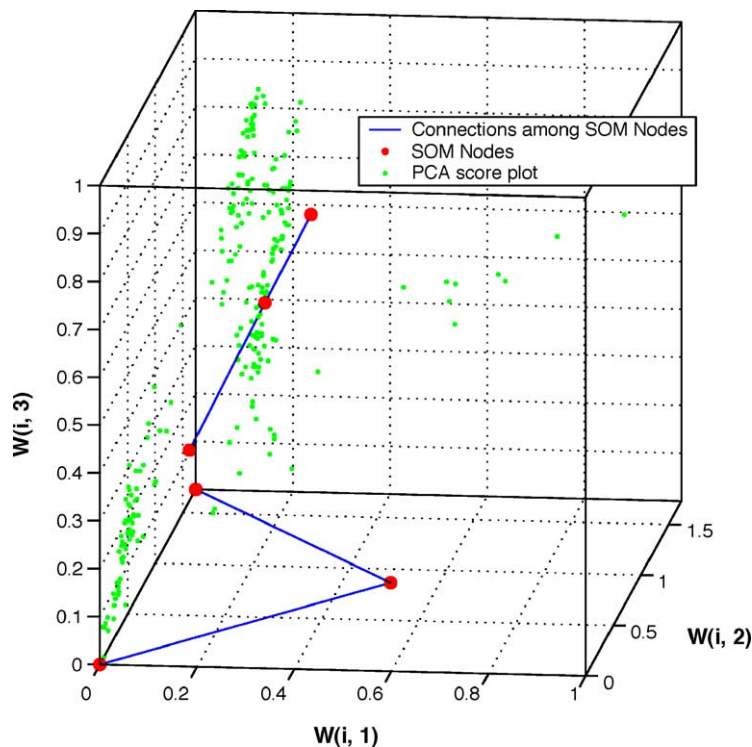

Fig. 6. Three-dimensional view of the unsupervised SOM analysis for three types of ENT bacteria data from polymer based e-nose.
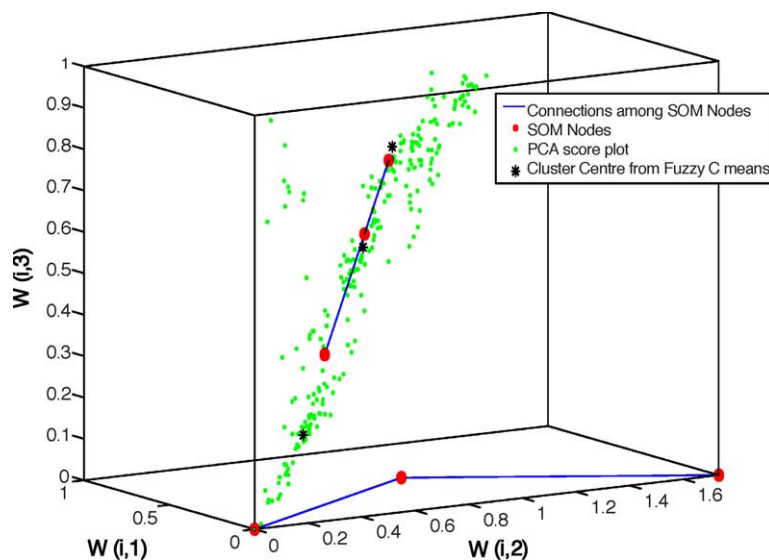
Fig. 7. Combined plot for the three types of ENT bacteria data, based on the extracted knowledge from the main data processing techniques, PCA, FCM and SOM, for the ENT bacteria class.

were plotted in multisensor space. So combining the three-dimensional scatters plot and FCM, cluster centres were properly located in multisensor space and also within the data. Various SOM networks were created and trained with the entire data set; a [3 × 1] SOM network performed best from all other SOM networks.

The objective of this analysis was to establish simple classes for the different bacteria species in order to examine if the data clusters could be separated for the conventional pattern recognition stage.

10. Evaluation of neural network—classification performance

The three different types of features of bacteria were analyzed using three supervised ANN classifiers, namely the multi-layer perceptron (MLP), probabilistic neural network (PNN) and radial basis function network (RBF) paradigms. Training of the neural networks was performed with 30% of the whole feature set. The remaining 70% of the whole feature set was used for testing the neural networks. These percentages were selected arbitrarily and were applied for all data sets. The aim of this comparative study was to identify the most appropriate ANN paradigm, which can be trained with best accuracy, to predict the "type of ENT infections" or in other words "type of ENT bacteria" [3,7].

10.1. Performance of MLP, RBF and PNN

10.1.1. Performance of MLP
A MLP network (with learning rate equal to 0.2 and a momentum term equal to 0.3) with 3–32 inputs and 3 output neurons was able to reach a success rate 78% in classification.

10.1.2. Performance of RBF and PNN
Neurons are added to the network until the sum-squared error (SSE) falls beneath an error goal (0.000001) or a maximum number (50) of internal neurons was reached. It is important that the spread parameter be large enough so that the radial basis neurons respond to overlapping regions of the input space, but not so large that all the neurons respond in essentially the same manner. For both the networks the spread parameter was set to 1.0. PNN was able to correctly classify 96% of the response vectors where as the RBF network's level of correct classification was up to 99.69%.

10.2. t-Test

A t-test was performed to assess if RBF, PNN were performing significantly better than the MLP in terms of the total number of patterns correctly classified. The null hypothesis H_0 demonstrated that there was no significant difference between the mean number of patterns misclassified by the RBF and PNN. The hypothesis H_0 was rejected at the 5% significance level ($t = 2.19$ for RBF and $t = 4.39$ for PNN).

11. Conclusion

This type of bacteria data analysis and feature extraction is very difficult. We can conclude that this combined use of three nonlinear methods (PCA based three-dimensional-scatter plot, SOM, FCM) can solve the feature extraction problem with very complex data and enhance the performance of Cyranose 320. Later on two supervised ANN classifiers, PNN and RBF were able to predict the three different ENT bacteria classes with 96 and 99.69% accuracy, respectively.

So from these results we can conclude that in future we can create a 'knowledge base of extracted features' by applying three nonlinear methods like PCA based three-dimensional-scatter plot, SOM and FCM for each bacteria class. So in future if we have an input dataset from unknown bacteria, by applying these three methods in a combined manner we can extract some feature for that unknown class of bacteria; later on we can match with the existing knowledge base of classes of bacteria features to predict the bacteria class. For this matching purpose, supervised ANN classifiers like PNN or RBF can be used with very high accuracy. Finally, we can conclude that this combined use of three nonlinear methods along with RBF neural network can solve the feature extraction problem with very complex data and enhance the performance of Cyranose 320 [11,12].

References

[1] http://www.show.scot.nhs.uk.
[2] http://www.ccm.lsuhsc-s.edu/bugbytes/Volume1/bb-v1n8.htm.
[3] T.C. Pearce, S.S. Schiffman, H.T. Nagle, J.W. Gardner (Eds.), Handbook of Machine Olfaction: Electronic Nose Technology, 1st ed., Wiley/VCH, 2003, ISBN 3-527-30358-8.
[4] http://www.cyranosciences.com.
[5] R. Dutta, E.L. Hines, J.W. Gardner, P. Boilot, Biomed. Eng. Online 1 (2002) 4.
[6] J.W. Gardner, M. Craven, C.S. Dow, E.L. Hines, Meas. Sci. Technol. 9 (1998) 120.
[7] http://www.mathworks.com.
[8] H.W. Shin, E. Llobet, J.W. Gardner, E.L. Hines, C.S. Dow, IEE Proc.: Sci. Meas. Technol. 147 (2000) 158.
[9] J.S.R. Jang, C.T. Sun, E. Mizutani, Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice-Hall, Upper Saddle River, NJ, 1997.
[10] T. Kohonen, Self-organising and Associative Memory, 2nd ed., Springer-Verlag, Berlin, 1987.
[11] J.W. Gardner, P.N. Bartlett, Electronic Noses: Principles and Applications, Oxford University Press, UK, 1999.
[12] J.W. Gardner, Sen. Actuators B 4 (1991) 108.

Biographies

Dr. Ritaban Dutta is graduated from University of Warwick, United Kingdom, with a PhD in electronics, for his PhD thesis "A novel approach to analyzing electronic nose data: object oriented expert system".

Presently he holds a Project Engineer position in "Environmental Condition Recognition Project" in International Automotive Research Centre, Warwick University and also working closely with http://www.medicalmicrosystems.co.uk/index.asp. He is also an Associate Fellow to the School of Engineering, Warwick University. In 2003, he has modelled a novel object oriented expert electronic nose system architecture which is based on object oriented technology and frame based expert system. This system architecture has been successfully simulated and evaluated with different electronic nose applications. This new design of electronic nose system on the basis of object oriented technology is his main contribution to the knowledge and core work for his doctorate research.

Dr. David Morgan, BSc, MB ChB, FRCS, FRCS Otol. is a Consultant Otolaryngologist and Skull Base Surgeon at Birmingham Heartlands and the Professorial Unit in Neurosurgery at University Hospital Trust. He qualified at Birmingham Medical School in 1981 and undertook surgical training in London, Boston, Sydney and Bombay. He is an Honorary Senior Lecturer and is currently a Surgical Specialty Advisor to the National Patient Safety Agency, a faculty member on regional and national otology courses and a medico legal advisor to the NHS Executive. In 1999, he co-founded Medicdirect Ltd, a health information website. In 2003, he formed Intelligent Medical Microsystems Ltd was formed following a 10 year University based research programme into the medical applications of expert computer systems. The private company is embedded in Birmingham Heartlands Hospital NHS Trust and was the first company to develop a tagging system to 'mistake proof' medical processes. He is also a main board member of Medilink – West Midlands and on the AIM health special interest group as well as NPfIT local implementation group.