**Enhancing e-nose performance by data fusion and sensor selection**

Julian W. Gardner[1], Pascal Boilot[2], Evor L. Hines[1]

School of Engineering, Warwick University, Coventry CV4 7AL, UK
[2] Current address, Alpha MOS, France

## Problem of high dimensionality

The human nose possesses around 100 million olfactory receptor cells following by a smaller, but still large, number of glomeruli nodes, mitral cells and tufted cells. We have around 300 distinct genes that encode olfactory receptor proteins and hence improve the specificity of olfaction [1]. The architecture of the human nose is thus one that poses a significant problem for most classification techniques – this is known as the "curse of high dimensionality". Even for more modest artificial noses with an array of just 32 sensors, the minimum number of features is 32 although it can be much higher when using dynamical information.

Figure 1 illustrates, qualitatively, the affect of a large number of features (e.g. sensors) on the successful classification of a pattern recognition problem. Having a large number of features tends to reduce the performance of the classification technique. For example, the larger number of sensors in an e-nose array, then the larger number of weights to learn in, say, a multilayer perceptron neural network. A fully connected array of 32 sensors with 16 neurons in the first processing and classifying just 10 different odours in the second output layer (the human system is around 8,000) would need over 670 weights to be learnt. Ideally, there would be replicated training vectors and so hundreds (if not thousands) of samples should be taken in order to solve the classification problem. Of course, there are "bootstrapping" techniques to train on smaller sets but the problem of high dimensionality still remains.

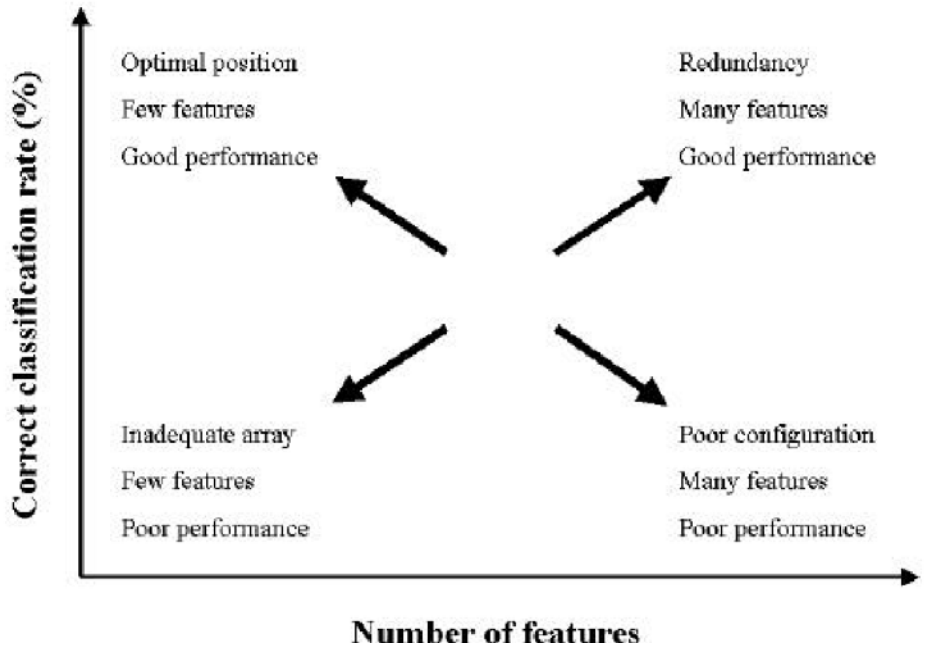


**Figure 1.** Configuration performance plot for feature selection [2].

The problem thus becomes one essentially of feature selection. In other words, if the number of features can be reduced to a more modest amount then the performance should improve for a given number of training vectors. The selection process removes the set of redundant features/sensors that simply add noise in to the system rather than improve discrimination. We believe that it is the process of inhibiting (i.e. eliminating) sensors is the key to solving a complex olfactory problem – while starting off with as large an array of sensors as possible.

**Problem of feature selection**

The classification problem can thus be improved by the identification of the optimal set of features within a larger set to solve a particular problem. For the moment let us consider that the feature is a sensor and so the problem is one of choosing the optimal set of sensors from an e-nose array. Now let is consider that we have an array of $n$ different sensors and determine the number of different permutations $p$ of possible sensors - ignoring $p$=0,1 as real options. The number of possible combinations $p$ (including using all sensors, $p$=$n$) is given by the equation

$$p = 2^{(n-1)} - 1 \qquad (1)$$

For an electronic nose with 32 sensors, the number of combinations $p$ is $2^{31}$-1, which is more than the number of water droplets in the Atlantic Ocean! Yet 32 sensors are well short of the human olfactory system which is known to have about 300 receptor proteins within the millions of receptor cells in total. We have also assumed here that the order in which the features are selected is not important, i.e. that the problem is a quasi static one. In the olfactory system, the order in which the cells are fired up may well be significant.

Thus, the problem of sensor selection is a critical one as the number of features (e.g. sensors) increases in an electronic nose. Feature selection is well know to researchers in this field and reports have been made of the use of various methods, such as linear projecting using PCA and LDA [2] and sequential search algorithms, such as the sequential forward search method (SFS) and sequential backward search (SBS) [3]. However, the problem with using a linear search method on a non-linear problem is clearly an issue.

**Fusion and sensor selection**

Here our approach is a combination of a fused sensor system (where possible) to increase the potential for odour discrimination but reduction in the dimensionality by a sensor selection technique based on genetic algorithms (GAs) [4]. GAs are attractive in that they can search through large sets of sensors in a relatively fast time and do so in a non-linear manner. Figure 2 illustrates the use of a v-integer genes GA selection technique to identify the best set of sensors [2]. The features are represented by the genes and these make up the chromosome. Once feature sensor has taken place, then a classification technique is required on that reduced set that should improve the overall success rate. Here we have used various methods to determine the value of fusion-selection, such as cluster analysis, fuzzy c-means, learning vector quantification, back propagation multiplayer perceptrons, radial basis function and probabilistic neural networks.

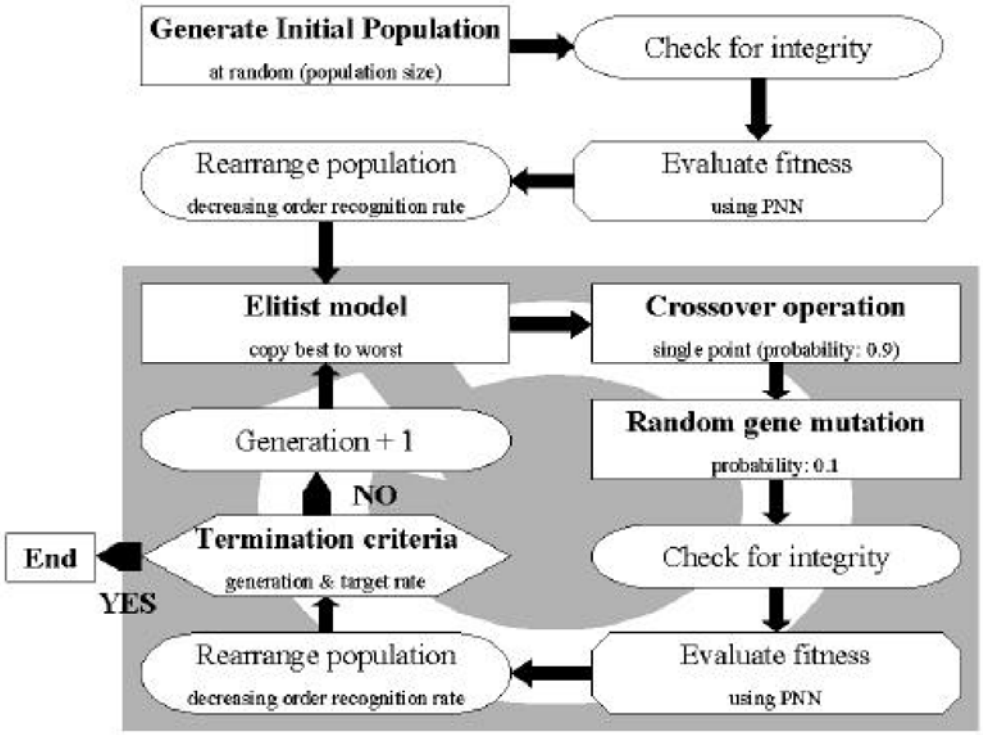


**Figure 2**. Flow diagram showing the use of a v-integer genes GA to solve the sensor selection problem [2].

## Data analysis and results

The algorithms have been tested on three principal data sets. Firstly, a set of standard fruit solutions with four different e-noses under an EU project. Secondly, two sets of bacterial samples – one eye and the other ear nose and throat infections using a Cyranose 320 (32 sensors). Thirdly, a set of banana samples with different ripeness levels using a metal oxide based array of just 4 sensors. For the sake of expediency, we concentrate here on the bacteria data sets.

SFS and SBS techniques were used to provide an initial exploration of the data but they only search a small percentage of the set of configurations unlike GA selection. However they did indicate that often just 6 sensors could produce a high classification rate on the bacterial data.

Table 1 shows the feature selection results on the eye bacteria data with a PNN classifier. The v-integer genes GA managed to find the optimal subset after 5 runs and the best subset of [8,11,15,23,31,31] achieved an impressive 90.6% success compared with 91.7% for the full set of 32 sensors.

Table 2 shows the results of different classifiers with the best set of six and three sensors. Again rather impressively, the MLP outperform the PNN classifier and show that with only 3 sensors out of 32 it is possible to achieve 93.3% success.

**Table 1:** Feature selection results for an e-nose array of 32 sensors with PNN classifier analysing eye bacteria.

| V-integer (No. of sensors) | Population (No. chromo.) | Random (Avg. init pop.) | GA Best % of all | GA Avg. % |
|---|---|---|---|---|
| 12 | 12 | 83.7 | 90.6 | 90.4 |
| 10 | 15 | 82.0 | 90.6 | 90.2 |
| 8 | 20 | 78.6 | 90.0 | 89.4 |
| 6 | 25 | 75.6 | 90.6 | 89.4 |
| 4 | 40 | 69.2 | 89.4 | 87.8 |

**Table 2:** Classification results for different techniques with the optimal subset of 3 and 6 sensors out of 32 for eye bacteria. (lr=learning rate).

| No. of sensors | Selected sensors | CA with 14 gps | FCM | MLP BPGDM (lr=0.1) | MLPBP LevMar | RBF sc=5 | PNN sc=0.05 |
|---|---|---|---|---|---|---|---|
| 6 | 8,11,15, 23,31,32 | 65% | 90% (16 clusters) | 87.8% | 96.7% (6×8×6) | 70.6% | 92.2% |
| 3 | 8,11,23 | 50.5% | 88.3% (13 clusters) | 90.0% | 93.3% (3×6×6) | 65.0% | 90.6% |

## Conclusions

For the generic solution to an e-nose problem, it is desirable to have a large number of features (e.g. sensors) such as in the human olfactory system. Consequently the fusion of e-nose systems or making larger arrays is an attractive proposition. However, the problems associated with high dimensionally make such systems impractical. Here we propose the use of a genetic algorithm based search algorithm to find the optimal subset of features (sensors), which then yield excellent results with a suitable predictive classifier. Once the subset has been identified for a particular e-nose problem, the classification problem is greatly reduced and removes the need for large number of training/calibration sets. The removal of most of the sensors/features in an array greatly reduces the level systematic drift/noise introduced by a large number of redundant sensors. Consequently, a GA sensor selection technique is a promising candidate for enhancing existing e-nose performance.

## References


1. T.C. Pearce, S.S. Schiffman, H.T. Nagle, J.W. Gardner (eds.) *Handbook of Machine Olfaction*, Wiley-VCH, 2003, 592pp.
2. P. Boilot, *PhD thesis*, University of Warwick, in preparation.
3. S. Pardo et al., in *Electronic Noses and Olfaction 2000* (eds. JW Gardner and KC Persaud), IOP Publishing, 2000, 83-88.
4. D.W. Aha and R.L. Bankert, in *Learning from Data* (eds. D Fisher and HJ Lenz), Springer-Verlag, New York, 1996, 199-206.