Available online at www.sciencedirect.com
SCIENCE DIRECT®
Sensors and Actuators B 106 (2005) 114–121
SENSORS AND ACTUATORS B CHEMICAL
www.elsevier.com/locate/snb

# Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach

J.W. Gardner*, P. Boilot, E.L. Hines

*Electrical and Electronic Engineering Division, School of Engineering, University of Warwick, Coventry CV4 7AL, UK*

Available online 23 July 2004

**Abstract**

Feature selection techniques can be used in order to find an optimal subset of sensors from an array of high dimensionality by eliminating redundant or irrelevant ones. By optimising the array size, the overall system performance can potentially be increased by maximising the information content and hence increasing the predictive accuracy. However, searching high dimensional space is problematic in the very high number of permutations. A novel search method procedure, *V-integer genes genetic algorithms (GA)*, is introduced and compared with other search methods such as sequential forward or backward searches (SFS or SBS) and *X-binary genes GAs*. Results are presented for a data-set consisting of 180 samples from some eye bacteria screening tests that were collected using an electronic nose (EN) with 32 sensing elements. For the data-set used in this work, SFS achieved over 89% correct classification by selecting just three features, whereas SBS needed at least five features to reach the same level. With *32-binary genes GAs*, the dimensionality is reduced by 50–60% and the classification rates are on average 91%. Considering eight, six or four features, the optimal subsets returned by the *V-integer genes GA* selections have dimensionality reduced by over 80% and on average achieve around 90% correct classification. Two selections, of six and three features, are considered for further pattern recognition (PARC) analysis using different classifiers. These results show that the newly developed *V-integer genes GA* approach is an accurate, and importantly, a very fast search method when compared to some other feature selection techniques.


*Keywords:* Feature selection; Genetic algorithm; Electronic nose; Array configuration; Search method; Sequential search; Integer; Binary

## 1. Introduction

The human nose possesses about 100 million olfactory receptor cells followed by a smaller, but still large number of glomeruli nodes, mitral cells and tufted cells. There are around 300 distinct genes that encode olfactory receptor binding proteins and hence improve the specificity of olfaction [1]. Electronic noses (ENs) are instruments that have been developed to mimic the 'human organ for smelling' [2]. An EN is both a chemical sensing and a data analysis system that can to some extent discriminate between different odours. Recent advances in the field of ENs has led to new developments in both sensors and feature extraction (pre-processing) and data processing techniques. As a result, the user of EN systems is provided with an increased amount of information (through the use of numerous methods) for the discrimination of odours using multi-sensor arrays. Even with a modest EN with an array of just 32 sensors, the minimum number of extracted features is 32 although it can be much higher when using dynamical information. This wide spectrum of sensors and features is desirable but not all of them are relevant to the pattern recognition (PARC) classification task. This recurrent problem when processing EN data is known as the *curse of high dimensionality* and has to be solved in order to try to optimise the performance of the system. The problem thus becomes essentially one of feature selection. The feature selection process removes the set of redundant features/sensors that are potentially adding noise into the system rather than improving discrimination. We believe that it is the process of inhibiting (i.e. eliminating) sensors which is one of the keys to solving a complex olfactory problem.

In this paper, the recurrent problem with EN of processing high-dimensional and redundant data-sets is first introduced. Section 2 introduces ideas relating to the possibility of finding an optimal array configuration of reduced dimensionality by considering a subset of features. It discusses different methods for feature selection and dimensionality reduction. The development of a new genetic algorithm (GA)

* Corresponding author. Tel.: +44 24 76 523695; fax: +44 24 76 418922.
*E-mail address:* j.w.gardner@warwick.ac.uk (J.W. Gardner).
*URL:* http://www.eng.warwick.ac.uk/SRL.

0925-4005/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2004.05.043

approach using integer coding instead of binary coding as an optimisation search method is introduced in Section 3. In Section 4, the data-set collected using the Cyranose 320 (Cyrano Sciences Inc.[1]) from samples in the eye bacteria screening tests, already presented in [3], is used to evaluate the techniques. Various feature selection techniques are used with this data-set to find an optimal subset of features. A number of non-linear PARC techniques are used to process this data by considering first of all 32 features and then reduced subsets of six and three features. Conclusions and discussions are presented in Section 5.

## 2. Feature selection

### 2.1. Problem of high dimensionality

For each type of gas sensor, tens of different sensors can be found which produce only slightly different responses to various volatile compounds. This wide variety of available gas sensors potentially creates a selection problem that is even more important in view of the cost and technology limitations. A reduction in the number of sensors is advantageous in the following circumstances:

- Sensors that are not sensitive to the target volatile compounds increase variance (noise) and do not assist with the smell recognition task.
- Sensors that have very similar sensitivities to the target volatile compounds provide redundant information (over complexity) not useful for the discrimination process.
- Additional sensors increase the weight, size and cost of a system (commercial issues) so that their number should ideally be minimised; without compromising the performance.

The process of sensor selection for a specific application is a trade off between factors such as system performance, cost and size [4]. When classifying multi-dimensional data, it is difficult to determine if all features (extracted from sensor responses) considered are necessary for the classifier. The introduction of irrelevant features increases the dimensionality of the search space, which can potentially overwhelm the accuracy of the pattern recognition (PARC) techniques. Most PARC classifiers can provide better recognition rate using a subset of features rather than the whole set, especially considering that irrelevant or redundant dimensions may be used [5]. The application can be simplified using smaller arrays with a reduced number of sensors and simpler classifiers working on fewer dimensions. Hence, a systematic or structured method is needed for selecting the optimum subset of sensors (optimal array configuration or key feature composition) and thus enhance overall system performance.

[1] Cyrano Sciences Inc. (USA), http://www.cyranosceinces.com.

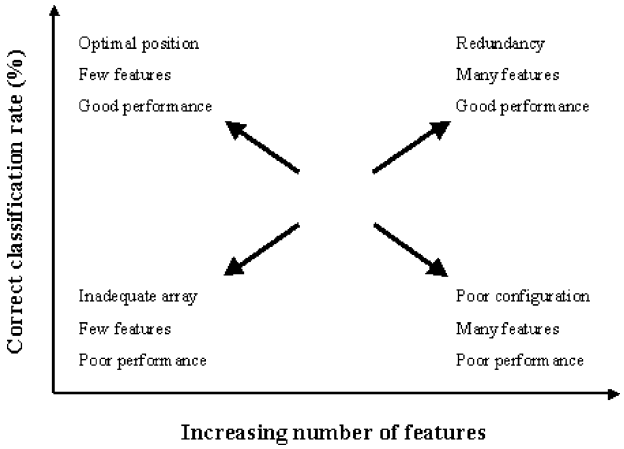


Fig. 1. Configuration performance plot for feature selection.

In this optimisation process, a measure of the correct classification rate can be used to maximise the performance of the system, whilst minimising the number of features selected [6]. Reducing the number of features in order to maximise the performance is achieved at certain cost, and features should be selected carefully as shown in the *configuration performance plot* in Fig. 1. This figure illustrates, qualitatively, the effect of a large number of features (e.g. sensors) on the successful classification of a PARC classifier. Too few, inappropriate or poor configuration of features can lead to an inadequate array performance. Conversely, increasing the array size may not be effective if the features selected are redundant and may also even reduce the discriminatory power of the system.

Having a large number of features tends to reduce the performance of the classification technique. Consequently, the number of training examples must be increased sharply with the number of features. This is in order to produce accurate results and to learn precise models. For example, the larger the number of sensors in an EN array, then the larger number of weights to learn in, say, a multi-layer perceptron (MLP) artificial neural network (ANN). A fully connected array of 32 sensors with 16 neurons in the first processing layer, and classifying just 10 different odours in the second output layer (the human system is around 8000) 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 produce accurate results and to solve the classification problem.

With EN data-sets, there are only usually a limited number of patterns (i.e. samples) available, there is an optimal number of features beyond which the performance of the PARC classifier starts to degrade. Problems with redundancy in data-sets, referred to as collinearity or correlation between features, are often due to the cross-selectivity of gas sensors. Hence, a structured approach for selecting the optimal sensors or features and reducing the dimensionality of the data matrix is typically required for a given classification problem [7]. This feature selection technique should

remove variables that introduce noise and those which are not representative. This should not only result in an increase in system performance and simplicity, but also potentially decrease the costs associated with data acquisition and the time needed to solve the classification problem. Most of the strategies reported in the literature for this task require the specification of an objective function and a search algorithm, as presented in the next section.

### 2.2. Problem of sensor selection

Various techniques have been presented in the literature that determine an optimal array configuration of reduced dimensionality by considering a subset of features. The simplest approach consists of evaluating each feature individually and selecting those with the highest scores, unfortunately, this approach ignores redundancy and will rarely find an optimal subset. Linear multivariate statistical techniques are often used to identify optimal discriminatory features and to help in defining the optimal array configuration; principal component analysis (PCA) or cluster analysis can be used to achieve this goal [8]. PCA is in practice an excellent technique to show poor sensor stability and sensor noise. The values (scores and loadings) derived using PCA have often been used to ascertain which features are the most discriminatory by evaluating the collinearities present in the sensor array and identifying the sensors with duplicate responses [9].

An exhaustive search of all combinations is often too computationally expensive since the number of all possible configurations for an array of $n$ elements grows rapidly. For the moment, let us assume that each feature is a sensor and so the problem is one of choosing the optimal set of sensors from an EN array. Now let us consider that we have an array of $n$ different sensors and that we wish to determine the number of different permutations $N$ of $p$ 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 Eq. (1).

$$N = \sum_{p=2}^{n} \frac{n!}{(n-p)!p!} \quad n \geq p \tag{1}$$

For an EN with 32 sensors, the number of combinations $p$ is many millions. Yet 32 sensors is far fewer than the human olfactory system which is known to have about 300 receptor proteins within the millions of receptor cells in total. It is 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 EN. 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 sequential search algorithms, such as the sequential forward search method (SFS) and sequential backward search (SBS) [5].

Sequential search algorithms, such as SFS and SBS, are greedy strategies that reduce the number of combinations to be tested by applying a local search [10]. With SFS, the selection algorithm starts with an empty space and a feature is added if it provides the best improvement of performance. With SBS, the algorithm begins from the whole feature space and a feature is extracted if it provides the least reduction in performance. SFS and SBS will only explore a small fraction of the whole set of configurations and can become trapped in local minima. Moreover, the problem with using a linear search method on a non-linear problem is clearly an important issue.

Randomised search algorithms, such as GAs, are inspired by the process of natural selection and perform a global random search on a population of solutions [11]. It is argued by some that GAs produce selections of features that are artefacts of the data with no physical insight, even so they are generic and always find a solution that is the best in some way. Pardo et al. [5] state that GAs are especially suited to this kind of selection problem, where every features can be easily coded in the chromosomes: a binary string where a one means the feature is present and a zero means the feature is absent. They follow the same procedure as the one initiated by Corcoran et al. [6,7] for which the multi-sensor array configuration is encoded as a chromosome, where each gene (location) has an allele (value) representing the presence or the absence of a sensor parameter. These and other issues will be discussed in the following section.

## 3. Integer-based genetic algorithm for feature selection

GAs are stochastic algorithms used to solve optimisation problems by working on a population of possible solutions. Starting from an initial population, a new population is created following a series of steps; selection, reproduction, mutation and competition [12]. Before applying GAs to an optimisation problem, one has to choose a *fitness function* (objective function) and to define a *suitable coding*. In most cases, when GAs are applied to feature selection problems, these are defined as follows:

- *Fitness function*. The basis for analysing an optimisation problem is to define a suitable fitness function, which will make it possible to compare possible solutions. The absolute minimum or maximum of the fitness function is the parameter being optimised by the GA. In this work, a wrapper strategy is used in order to evaluate subset solutions using their predictive accuracy (based on the results for a given PARC technique).
- *Suitable coding*. One of the most important task in problem solving using GAs is coding the solution space, and the best-known forms of coding are strictly binary. Any potential solution is coded as a vector called a chromo-

some, the elements of which are called genes and are located in well-defined positions (alleles). The chromosomes representing the entire set of features can easily be coded as a binary string where 1 means the feature is present and 0 means the feature is absent [5,6]. For example, given eight sensors from which an optimal configuration needs to be identified, the chromosome 01010011 will represent an array configuration with sensors 2, 4, 7 and 8. A number of chromosomes are then used to form a population of possible solutions.

The selection algorithm uses a genetic representation for which each feature is equivalent to a gene; hence, the length of the chromosome is equal to the total number of features. This type of GA optimisation will be referred to as *X-binary genes GA*, where *X* is the number of features to choose from. In summary, the GA search process typically comprises of the following steps:

- Step 1. Generate initial population of chromosomes from the global feature set.
- Step 2. Evaluate fitness (objective function) of each chromosome.
- Step 3. Are the optimisation criteria met? If YES, go to step 7. If NO, go to step 4.
- Step 4. Generate new population by selecting pairs for mating, recombination using mask-based (or template-based) crossover and mutation.
- Step 5. Evaluate fitness (objective function) of each new chromosome.
- Step 6. Identify the fittest individual in the population. Go to step 3.
- Step 7. End.

As reported in the literature [5,6], this representation is the most suitable for many optimisation problems. For particular selection problems of large dimensions, the chromosomes formed might be too complex and too long to handle easily, thus considerably slowing down the optimisation process. Therefore, it was decided to apply a new and original coding technique for the creation of the population of chromosomes [13]. For the genetic representation, *V*-genes long chromosomes were used with integer values from one to the number of features, representing the selected subset of features. This type of GA optimisation will be referred to as *V-integer genes GA*, where *V* is an input parameter defined by the user representing the number of features to be selected.

Fig. 2 shows a schematic representation of the way this technique works [14]. The initial population is generated at random. The resulting classification rates at this stage are used for comparison to produce the results for randomly selected subsets. There are no rules for defining the size of the initial population, for *X-binary genes GAs* a population of 15–20 chromosomes will be considered. In order for *V-integer genes GAs* to converge, they must be given a large number of genes to work on so that the genetic operators can apply. The smaller the chromosomes, the larger the population. It was found experimentally that the initial population should contain at least three times as many genes as the total number of features. In order for this GA optimisation to be successful, one of the conditions is that there will be no repetition of one feature in the same chromosome, referred to as the *check for integrity* procedure. With every generation, the recognition rate (objective function) is evaluated using probabilistic neural network (PNN) classifiers for every individual chromosome within the population. PNN was selected because it converges a lot faster than the traditional multi-layer perceptron (MLP) network and works well with the data-sets considered here. Thanks to the objective function, the best selection subset of features achieving the best recognition rate is found.

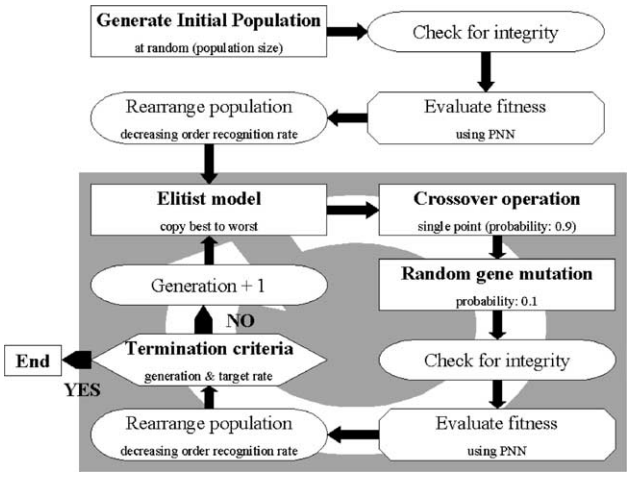


Fig. 2. Flow diagram showing the use of *V-integer genes GA* for feature selection.

The survival of the fittest (elitist model) was used to ensure that the best chromosome will survive to the next generation and will not be affected by the genetic operators. When the population is rearranged, chromosomes are sorted in ascending order of recognition rate. Then, a copy of the fittest is used to replace the worst before the genetic operators are applied. *For X-binary genes GAs*, the chromosomes are long enough and it was decided to apply a mask- or template-based crossover. Considering the relatively small length of the chromosomes, it was decided to apply single point crossover operations for *V-integer genes GAs*. In each population, two parents are selected at random and genes are crossed over at a random point to constitute the offspring chromosomes of the new generations. The probability of one chromosome being affected by a crossover operation was set quite high to 0.9 so that the GA explores the search space more rapidly. Based on experience, a random gene mutation stage was introduced with the probability of one gene being affected being set to 0.1 to significantly change the information in each new chromosome. After applying the genetic operators, the *check for integrity* is performed, and the fitness of each new chromosome is evaluated using the classification rate from the PNN results. Two optimisation

criteria were selected that will terminate the GA optimisation process.

- The maximum number of generations was set to 20 and represents the number of times the search algorithm is repeated.
- In order to determine that a valid solution to the problem has been found, the target recognition rate was set to over 95% in order to achieve the same level of performance as, or better than, the full set of features.

Both SFS and SBS are considered, together with *X-binary genes GA* and the new *V-integer genes GA* selection techniques. These various selection techniques were applied on the data-set collected using the Cyranose 320 (Cyrano Sciences Inc.[1]). A total of 180 readings is included in this data-set representing the three dilutions of six bacteria species [3]. PNN is used as the classifier for evaluating the fitness function for each chromosome solution. The results of this approach are presented in the following section.

## 4. Data analysis results

A Cyranose 320 (Cyrano Sciences Inc.[1]) electronic nose was used to sample eye bacteria, it comprises of a 32-sensor-array of carbon black polymer composite resistors. As described by Boilot et al. [3], PCA was first used to examine the data clustering in multi-sensor space and to verify that the clusters established were genuine and matched the six types of bacteria. Although these results show that the data can be easily linearly separated, they also show that the sensor responses are highly correlated as most of the variance in the data is contained within PC#1 (99.68%). Most of the sensors are cross-selective being sensitive to the same odorous constituents hence they become redundant for the classifier. Ideally, given this database of the responses of 32 sensors to six bacteria species, it should be possible to identify a subset of features that provide the same, or better, level of performance when applied to the classification of these pathogens. The aim of this application is to investigate how to best create an optimal, or near-optimal, array configuration using a reduced number of sensors from the 32 available. Selecting the most appropriate combination of sensors is needed in order to satisfy the needs of the current application in terms of optimising the solution cost and reliability of the system.

### 4.1. Sequential forward and backward search

Search methods were first used to guide the feature selection process, the correct classification rate is determined by PNN. SFS and SBS were implemented first because they offer the potential advantage of a step-by-step scanning of the whole feature space. With SFS, the selection algorithm starts from an empty space and the feature that provides the best improvement of the objective function is added to the selection. This process is repeated until all sensors have been selected. With SBS, the algorithm begins with the whole feature space and the sensor that provides the least reduction of the objective function is removed from the selection. This process is repeated until all sensors are discarded. After implementing these two search methods, it is possible to represent the evolution of the recognition rate against the number of sensors, as presented in Fig. 3.

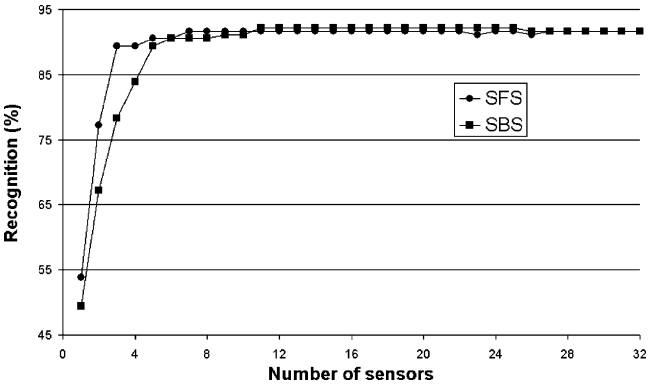


Fig. 3. Recognition rate for SBS and SFS applied to the eye bacteria data.

When the number of sensors selected exceeds five, the two curves flatten out reaching a maximum recognition rate of around 91%. As more sensors are added to the selection, the final recognition rate remains constant, these additional sensors are not necessary to maximise the predictive accuracy. SFS presents a faster response to reach this optimum state and when three sensors are selected, namely {8,11,23}, the recognition is already at 89.4%. With SBS removing one sensor from the selected five sensors {1,8,11,14,32} causes the recognition rate to drop below 89.4%. However, even though the solutions returned with SFS and SBS achieved relatively high classification rate, they only explore a small percentage of the whole set of configurations (the order in which sensors are added is also critical) and these solutions might not be optimal.

### 4.2. Binary and integer-based genetic algorithm selection

Both *X-binary genes GA* and *V-integer genes GA* feature selection techniques are considered here. The initial population for each GA approach was generated randomly. The classification rates found at this stage were used for comparison to produce the results for the randomly generated subsets. The recognition rate is evaluated in every generation using PNN for every individual solution within the population, as follows:

- *X-binary genes GA*. For the genetic representation, the chromosomes representing the entire set of features can easily be codified as a binary string where 1 means the feature is present and 0 means the feature is absent. This representation was adopted with the chromosomes representing the selected features being 32-genes long. There

Table 1
GA feature selection results for an EN of 32 sensors with PNN classifier

| Chromosome structure | Population size | Random (initial population) | | GA process (last population) | |
|---|---|---|---|---|---|
| | | Average | Best of all | Average | Best of all |
| 32-binary | 20 | 90.0 | 90.6 | 91.1 | 91.1 |
| 12-integer | 12 | 88.9 | 89.4 | 90.4 | 90.6 |
| 10-integer | 15 | 87.8 | 88.3 | 90.2 | 90.6 |
| 8-integer | 20 | 87.8 | 88.9 | 89.4 | 90.0 |
| 6-integer | 25 | 87.2 | 88.3 | 89.4 | 90.6 |
| 4-integer | 40 | 85.5 | 86.7 | 87.8 | 89.4 |

are no rules for defining the size of the initial population, a population of 20 chromosomes was considered for *32-binary genes GAs*.

- *V-integer genes GA*. For the genetic representation, *V-integer genes* long chromosomes were used with integer values from 1 to 32 representing the selected sensors. The number of chromosome in the initial population was set so that the total number of genes in the population (number of chromosomes × number of features) will be at least five times the total number of sensors, around 150. Various combinations were investigated such as; 12 chromosomes for *12-integer genes GAs*, 15 chromosomes for *10-ineger genes GAs* and 40 chromosomes for *4-integer genes GAs*.

For *V-integer genes GAs*, single crossover points are applied in each population and *for 32-binary genes GAs*, a mask is created at the start to define which genes should be swapped over. The number of crossovers to apply per generation is set so that the probability that one chromosome is affected is about 0.7. Random point gene mutations are then carried out on the chromosomes, randomly selecting a gene from the entire population and changing its value. The number of mutations to apply per generation is set so that the probability that one gene is affected is about 0.1. Two optimisation criteria were selected; the maximum number of generations was set to 20 and the target recognition rate was set to over 95%. Table 1 presents the results of the selection process after 20 generations in terms of recognition rate. The GA optimisation process was run five times, at the end of the process (after 20 generations), the results for the best solution (chromosome) is identified. In Table 1, the average of these fittest chromosomes over five runs is included as well as the result for the best solution found. At the start of the GA process, the first initial population is generated at random, the results for the best solutions found in this way is also included in this table in terms of the average over five runs and the best solution.

In general, the results at the end of the GA process are always better than for randomly generated solutions. The GA feature selection always finds a better solution representing an optimal array configuration. With *32-binary genes GAs*, examination of the complete set of results shows that the GA search method consistently identified subsets containing between 12 and 16 features. These subsets have optimal classification rates of around 91.1% (91.1%, at best) when compared with randomly generated solutions, which typically have classification rates of around 90.0% or below (with one exception at 90.6%). With this GA feature selection technique, the array dimensionality is reduced by 50–60% and some of the redundant sensors are identified. A large number of features are considered and this technique shows its limit for this data-set of large dimensionality.

With the *X-integer genes GAs*, as the number of features to select becomes smaller, the recognition rate of the subsets generated at random decreases from 88.9% with 12 features down to 85.5% with only four sensors. The *V-integer genes GA* results also follow this trend but more slowly, with fewer features, it becomes more difficult to find an optimal solution. *V-integer genes GAs* always manage to find an optimal subset of sensors so that the recognition rate of the best of all after five runs is nearly always 90.6%. With the full set of sensors, we achieved 91.7% correct classification with PNN (spread constant of 0.1). With *6-integer genes GAs*, the array dimensionality is reduced by 81%, the best subset found is {8,11,15,23,31,32} achieving 90.6%. The *4-integer genes GAs* (dimensionality reduced by 88%) could consistently identify subsets of sensors that will achieve high levels of performance: 87.8% on average and 89.4% at best with {8,11,23,32}.

### 4.3. Results when using 32, 6 and 3 sensors

In order to assess if the subsets formed performed as well or better than the full set of sensors, it was decided to run a series of tests using different classifiers. In the GA process, PNN was used but other classifiers might perform better and it was decided to investigate further the recognition rate of some of the selected subsets. The fittest six-sensor subset returned by the GA selection process {8,11,15,23,31,32} was considered because it was selected using a survival of the fittest strategy. A statistical study carried out on the subsets created showed that these six sensors were the most likely to be selected. It was decided to reduce the selection even further and to consider only three sensors based on their selection occurrence. The generated subset contains {8,11,23}; this is the same subset as the one returned using SFS that achieved 89.4% with PNN. Two clustering techniques were

Table 2
Classification results for different PARC techniques with sets of 32, 6 or 3 sensors

| Number of sensors | Selection | CA (%, groups) | FCM (%, clusters) | MLP BPGDM (%) | MLPBP LevMar (%) | RBF (%, spread) | PNN (%, spread) |
|---|---|---|---|---|---|---|---|
| 32 | All | 65.0 (14) | 90.6 (14) | 94.3 (32 × 18 × 6) | 97.3 (32 × 18 × 6) | 92.2 (5) | 91.7 (0.1) |
| 6 | [8,11,15,23,31,32] | 65.0 (14) | 90.0 (16) | 87.8 (6 × 8 × 6) | 96.7 (6 × 8 × 6) | 70.6 (5) | 92.2 (0.05) |
| 3 | [8,11,23] | 50.6 (14) | 88.3 (13) | 90.0 (3 × 6 × 6) | 93.3 (3 × 6 × 6) | 65.0 (5) | 90.6 (0.05) |

used, namely cluster analysis (CA) using nearest neighbour and fuzzy C-mean (FCM), together with four supervised artificial neural network (ANN) classification paradigms; MLP trained back-propagation using the gradient descent with momentum (BPGDM) or the Levenberg–Marquardt (BPLevMar) algorithm, radial basis function (RBF) and PNN. The results are presented in Table 2.

The classification rate results obtained with a subset of three sensors are less than with a subset of six sensors, higher predictive accuracy suggests better information content for the selection considered. For FCM, the results with the two subsets are similar (around 90%) and about the same levels are achieved with MLP BPGDM. The best results were obtained with MLP BPLevMar and PNN where the classifiers achieved, respectively, 96.7% and 92.2% with six sensors, and 93.3% and 90.6% with three sensors. These levels of performance are very similar to those obtained with the full set of parameters (slightly lower in general). Using *V-integer genes GAs*, a subset of six (or three) sensors is selected that will perform as well as the full 32-sensor-array in the discrimination of the bacteria of interest. This demonstrates the potential for developing an application-specific instrument insofar that an optimal array of sensors can be successfully created thus potentially reducing the cost of the system.

## 5. Conclusions

Feature selection techniques can be used in order to develop an application-specific instrument by eliminating redundant or irrelevant sensors, therefore, optimising the array configuration. The removal of most of the sensors/features in an array can greatly reduce the level of drift/noise introduced by a large number of redundant sensors. These results show that the newly developed *V-integer genes GA* approach is an accurate and fast search method when compared to the other feature selection techniques investigated here. The levels of performance in terms of predictive accuracy of the subsets selected using *V-integer genes GAs* are the same as (or better) than those achieved with the more popular and well-established *X-integer genes GAs*. *V-integer genes GAs* outperform *X-integer genes GAs* by finding subsets with fewer features and reducing further the dimensionality of the problem. This new GA-based approach will prove to be reliable and beneficial in defining an optimal array configuration for a particular application. Using this selection approach, one can find an optimal subset of sensors for a particular application while choosing sensing devices from a larger database of sensors. These techniques will help to create smaller application-specific sensor arrays and help to reduce the potential cost associated with new sensor developments.

In most of the applications and the commercial EN systems available, the broad and overlapping specificities of the sensors used (e.g. metal oxide semiconductors or conducting polymers) create an array for which the discrimination capability often exceeds the needs of the application (i.e. redundancy exists). The aim of sensor selection is to select an optimised (smaller) array of sensors (subset) for a particular application from a larger set of available sensors (library/database). Ideally, given a reference database of odours from $n$ sensors, it should be possible to identify a subset of $n'$ sensors that can produce the best possible discrimination of the different samples of interest [6]. For EN based on integrated arrays, such as the Cyranose 320, the cost associated with the design of a new array for a specific application will be significant. So in this particular case, it is best to keep the integrated sensor chip but switch a subset ($n$–$n$) 'off' for certain applications. The strategy for other commercial EN systems, e.g. manufacturers like Alpha M.O.S.[2] is to offer smaller and more cost effective solutions for which the sensors have been selected a priori for a particular application.

## Acknowledgements

We thank Micropathology Ltd. for the use of laboratory and assistance when running the tests, Cyrano Sciences Inc. (USA) for the EN system and technical assistance. We also thank Ross Folland and Mario Gongora (Warwick University) for their active collaboration and helpful support in processing the data. Pascal Boilot gratefully acknowledges financial support (award number 99310943) from EPSRC during his time as a Ph.D. student at the University of Warwick.

## References

[1] T.C. Pearce, S.S. Schiffman, H.T. Nagle, J.W. Gardner (Eds.), Handbook of Machine Olfaction, Wiley–VCH, 2003.

[2] Alpha M.O.S. (France), http://www.alpha-mos.com.

[2] J.W. Gardner, P.N. Bartlett, A brief history of the electronic nose, Sens. Actuators B 18–19 (1994) 211–220.
[3] P. Boilot, E.L. Hines, J.W. Gardner, R. Pitt, S. John, J. Mitchell, D.W. Morgan, Classification of bacteria responsible for ENT and eye infections using the Cyranose system, IEEE Sens. J. 2 (3) (2002) 247–253 (special issue on artificial olfaction).
[4] B.G. Kermani, S.S. Schiffman, H.T. Nagle, A novel method for reducing the dimensionality in a sensor array, IEEE Trans. Instrum. Meas. 47 (3) (1998) 728–741.
[5] S. Pardo, S. Marco, C. Calaza, A. Ortega, A. Perera, T. Sundic, J. Samitier, Methods for sensor selection in pattern recognition, in: J.W. Gardner, K.C. Persaud (Eds.), Electronic Noses and Olfaction 2000, IoP Publishing, 2000, pp. 83–88.
[6] P. Corcoran, J. Anglesea, M. Elshaw, The application of genetic algorithms to sensor parameter selection for multisensor array configuration, Sens. Actuators A 76 (1999) 57–66.
[7] P. Corcoran, P. Lowery, J. Anglesea, Optimal configuration of a thermal cycled gas sensor array with neural network pattern recognition, Sens. Actuators B 48 (1998) 448–455.
[8] A.D. Walmsley, S.J. Haswell, E. Metcalfe, Methodology for the selection of suitable sensors for incorporation into a gas sensor array, Anal. Chim. Acta 242 (1991) 31–36.
[9] J.W. Gardner, Detection of vapours and odours from a multi-sensor array using pattern recognition. Part 1: principal component analysis, Sens. Actuators B 4 (1991) 109–116.
[10] D.W. Aha, R.L. Bankert, A comparative evaluation of sequential feature selection algorithms, in: D. Fisher, H.-J. Lenz (Eds.), Learning from Data, Springer–Verlag, 1996, pp. 199–206.
[11] D.E. Goldberg, Genetic Algorithms in Search, Optimisation and Machine Learning, Addison-Wesley, 1989.
[12] M. Russo, L.C. Jain, An introduction to evolutionary computing, in: L.C. Jain (Ed.), Evolution of Engineering and Information Systems and Their Applications, CRC Press, 1999, pp. 1–30.
[13] P. Boilot, E.L. Hines, M.A. Gongora, R.S. Folland, Electronic noses inter-comparison, data fusion and sensor selection in discrimination of standard fruit solutions, Sens. Actuators B 88 (2003) 80–88.
[14] P. Boilot, Ph.D. Thesis, University of Warwick, July 2003.