Improved chemical identification from sensor arrays using intelligent algorithms

Thaddeus A. Roppel*[a], Denise Wilson[b]

[a]Department of Electrical and Computer Engineering, Auburn University
[b]Department of Electrical Engineering, University of Washington

ABSTRACT

Intelligent signal processing algorithms are shown to improve identification rates significantly in chemical sensor arrays. This paper focuses on the use of independently derived sensor status information to modify the processing of sensor array data by using a fast, easily-implemented "best-match" approach to filling in missing sensor data. Most fault conditions of interest (e.g., stuck high, stuck low, sudden jumps, excess noise, etc.) can be detected relatively simply by adjunct data processing, or by on-board circuitry. The objective then is to devise, implement, and test methods for using this information to improve the identification rates in the presence of faulted sensors. In one typical example studied, utilizing separately derived, a-priori knowledge about the health of the sensors in the array improved the chemical identification rate by an artificial neural network from below 10 percent correct to over 99 percent correct. While this study focuses experimentally on chemical sensor arrays, the results are readily extensible to other types of sensor platforms.

Keywords: Intelligent signal processing algorithms, neural networks, sensor arrays, fault detection

1. INTRODUCTION

A common requirement in the application of chemical sensor arrays is the need to tolerate permanent or transient sensor faults. Under some circumstances this goal can be addressed by modifying existing signal processing algorithms or by incorporating redundant sensors into the design[1]. However, the latter approach can be problematic, since, for chemical sensors in particular, it can be difficult to maintain calibration for a standby sensor which is called into service only rarely. A variation of this scenario is that while a replacement sensor may be available, a finite time is required to bring it on-line. The approach described here is to identify a faulted sensor, remove the corresponding data stream, and synthesize a suitable replacement input data stream. We show that under certain commonly-encountered conditions it is possible to do this with minimal loss of accuracy as long as the universe of measurement does not deviate too far from the training data universe during the transition time.

Other methods for handling sensor faults have been described in the literature. Roppel et al. describe training a neural network on all possible single-sensor faults in a multi-sensor array[2]. In this case, redundant sensors are employed, so that the neural network essentially implements a voting scheme. The uniqueness of this approach is that the boundaries of acceptable performance for each sensor are determined by training rather than by preset values. This provides flexibility to accommodate a variety of sensor systems and applications. However, this approach becomes unwieldy for larger sensor arrays due to the extensive training required. Klein describes a sensor redundancy scheme based on Boolean algebra which has the attribute of functioning even without specifically identifying which sensor(s) are faulted[3]. Willett et al. present a detailed mathematical analysis of the results of employing like-sensor fusion, and are able to lay out the conditions under which maximum benefit is obtained[4].

2. EXPERIMENTAL METHOD

2.1 Sensor Testbed

In this study, the experimental testbed consists of an array of 4 surface acoustic wave (SAW) sensors, each with a different chemically sensitive coating. The array is exposed to 8 chemical vapors one at a time, and the array response is recorded as

* Correspondence: Email: troppel@eng.auburn.edu; http://www.eng.auburn.edu/~troppel

Advanced Environmental and Chemical Sensing Technology, Tuan Vo-Dinh, Stephanus Buettgenbach, Editors, Proceedings of SPIE Vol. 4205 (2001) © 2001 SPIE. · 0277-786X/00/$15.00

the concentration is varied over the range from 5% to 45% of saturation. The four sensor responses comprise the input vector to an artificial neural network (ANN). The raw data are values of phase shift derived from the SAW sensors. For analysis, the ANN inputs are normalized to lie in the range from 0 to 1.

For part of the experimental investigation, the data set is purposely modified so that one sensor appears faulted, while the remaining sensors are contaminated with uniform pseudo-random noise. It is presumed that the fault condition is determined by independent means.

The data are first viewed using two-dimensional principal components analysis (PCA) in order to discover general trends, and then the performance of a neural network is studied in detail. The goal of the experiment is to investigate the ability of the neural network to identify the eight analytes when the faulted sensor data are replaced by suitably chosen synthetic data. The performance is measured as a function of concentration, and as a function of the level of noise added to the sensors.

2.2 Principal Components Analysis (PCA)

PCA is widely used as a quantitative tool for measuring properties of data sets, and as a visualization technique to help understand relationships among data clusters within multi-dimensional data sets[5]. We employ two-dimensional PCA here for the latter application, and use the results to guide the training and testing of a neural network on the original data.

2.3 Artificial Neural Network (ANN) Architecture

The ANN employed in this investigation is a fully-connected multi-layer perceptron with one hidden layer. There are four inputs (one for each sensor), eight outputs (one for each analyte to be identified), and 12 hidden layer neurons. The training is performed using extended delta-bar-delta error back-propagation with initial learning coefficient of 0.3 and momentum of 0.4. The hidden layer neurons use a hyperbolic-tangent activation function. Simulation of the ANN was performed using NeuralWorks Professional II/Plus[6].

2.4 Data Synthesis

For the purpose of investigation, one sensor is considered to be faulted so that its value is unknown, and the corresponding ANN input is replaced by synthetic data. The synthetic data is obtained sweeping the unknown sensor input through its entire range (0 to 1) in discrete steps while holding the other three inputs fixed, and picking the value that yields the highest confidence at the ANN output. The confidence measure is defined in Section 4.1

3. INVESTIGATION OF THE DATA USING PCA

3.1 Analysis of the Raw Data Using PCA

Figure 1 shows a PCA plot of the data set before any modifications are artificially introduced. The eight analytes are labeled with the letters *a* through *h* on the plot. Each labeled point corresponds to a unique concentration. The plot shows that for the most part the data are "well-behaved," by which we mean that as the concentration of each analyte changes, the track in PC space is relatively smooth and continuous. For higher concentrations, the analytes are well separated in PC space, but there is significant overlap at lower concentrations. This suggests that the analytes might not be successfully distinguished at lower concentrations; this is confirmed by the ANN results.

We have selected to artificially fault one sensor (S-5) and to observe the effect of replacing it with a synthetic data vector. In order to see the contribution of S-5 to the overall data set, we plot the PCA results just using the other three sensors [2, 3, 7] in Fig. 2. The separation among analyte traces is poorer than in Fig. 1, confirming that S-5 does provide useful data.

3.2 PCA Including the Synthesized Data

Figure 3 shows a PCA plot which includes three examples of the best-match approach using synthesized data. Specifically, we imagine that the sensor array is presented with analyte "e" at maximum concentration, but sensor S-5 is known to be faulted. Then we generate the line labeled "test 1" by fixing the trained-on values of the three good sensors, and sweeping the synthetic input of sensor S-5 in 20 discrete steps over its range (0 to 1). One of these 20 steps results in a "best match" to the training data, as evidenced by the intersection of the line with the desired data point (the leftmost "e" on the track). Although we can identify which step results in the best match and thus the best replacement sensor value, it is more important that we have identified the analyte, which is our objective.

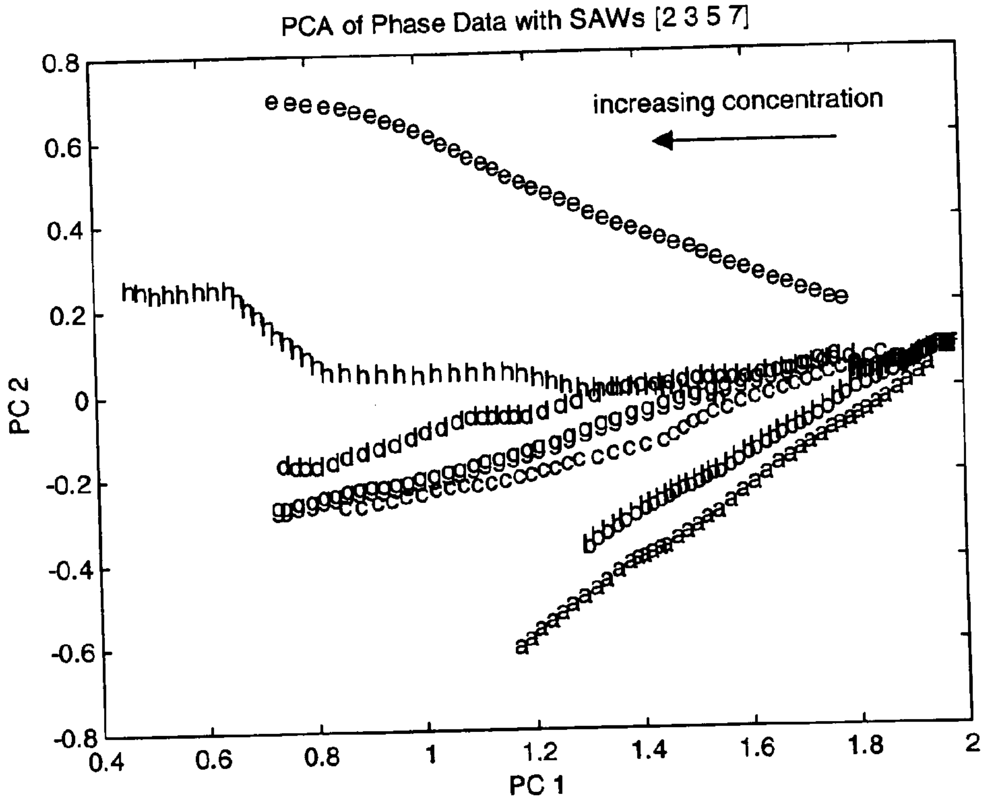

Figure 1. PCA plot of data from four SAW chemical sensors exposed to eight analytes labeled *a* through *h* at 41 concentrations ranging from 5% to 45% of saturation. The arrow shows the trend direction for increasing concentration.

Furthermore, we have even identified the concentration of the analyte, although this was not a goal we set out specifically to achieve. The other two test lines show similar results for other analytes at other concentrations. The line labeled "test 2" corresponds to analyte "c" at minimum concentration, and line "test 3" corresponds to analyte "a" at its 39th concentration step, which is the third labeled point from the left.

The PCA results illustrate that the "best-match" approach can identify at least some analytes with high confidence and selectivity. This provides motivation for a more quantitative investigation using an ANN.

4. NEURAL NETWORK RESULTS

For quantitative studies, an artificial neural network is trained on good sensor data, and then tested on inputs in which one "bad" sensor input is replaced by a synthesized vector of hypothesis values while the good inputs are frozen. In this case, the "best match" is determined by a quantifiable criterion applied to the neural network outputs. Namely, the hypothesis value that results in the most confident output is accepted as correct. Then this determination is compared to the truth. Results in the form of confusion matrices are averaged over concentration at various noise levels, and also examined as a function of concentration.

4.1 Definition of Confidence Level

In order for the "best-match" approach to work, it is necessary to devise an algorithm to calculate a confidence level associated with the output for any given input. It is desirable for this algorithm to be as simple as possible to compute, while yielding good results. The algorithm we employ here is to calculate the ratio of the maximum element of the output vector to the average of all the elements.

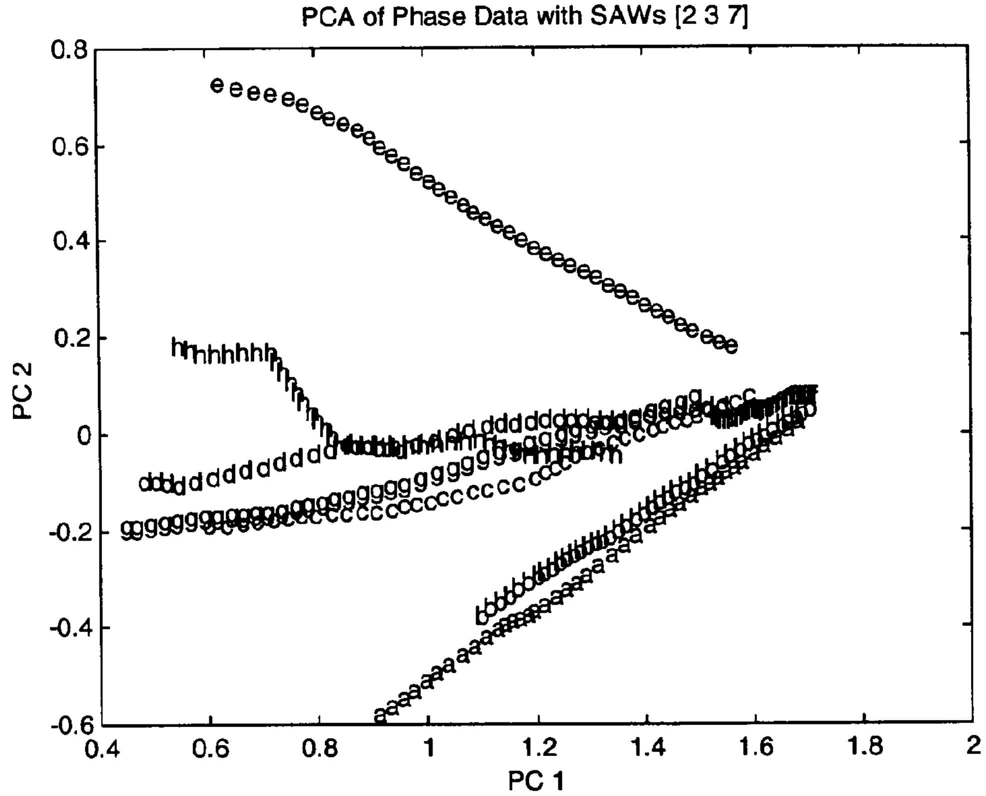

Figure 2. When one of the four sensors (S-5) used in Fig. 1 is left out, the separation among analyte tracks becomes poorer, indicating that S-5 contributes significantly to the data set.

In this way, an "indecisive" output vector in which all elements have roughly the same value will receive a low score, while a "confident" output vector in which one element is large and the rest small will receive a high score. The simplicity of this algorithm renders it attractive, and it should be possible to implement it relatively simply in a practical application.

4.2 Test Results Without Noise

Table 1 shows the confusion matrices averaged over all concentrations for the testing data and the training data. A single-valued numerical measure of the overall identification accuracy can be calculated as the average of the confusion matrix elements along the major diagonal. Henceforth we will call this the CMDA. Using this metric, the accuracy for the testing data is 75.6%, while the training data averages 96.6%. Although the confusion matrix is not shown here, we also calculated the accuracy for a worst-case scenario in which the faulted sensor is not replaced, but is stuck at zero. The CMDA in this case is less than 10%.

As predicted by the PCA results, the ANN performance is a function of concentration. Table 2 lists the values of the CMDA metric broken out by concentration range. The results show that performance ranges from 56.8% at very low concentrations to 87.5% in the top half of the concentration range. Furthermore, if analytes *g* and *h* are combined as a detection class, the CMDA is 100% in the higher concentration regime.

4.3 Test Results With Simulated Noise Added to the Sensor Responses

Noise (uniformly distributed, pseudo-random) was added to the sensor responses at three different levels: 1%, 5%, and 10% of full scale range. Table 3 shows the CMDA metric for the testing and training data at these noise levels. These data show that adding noise degrades the ANN performance for both the training data and the testing data; however the impact on the testing data is more severe, especially at lower concentrations. In the highest concentration range (31 – 41), the ANN yields 100% CMDA for 1% noise. For the lowest concentration range (1 – 10) with 10% noise, the CMDA drops to 46.5%.

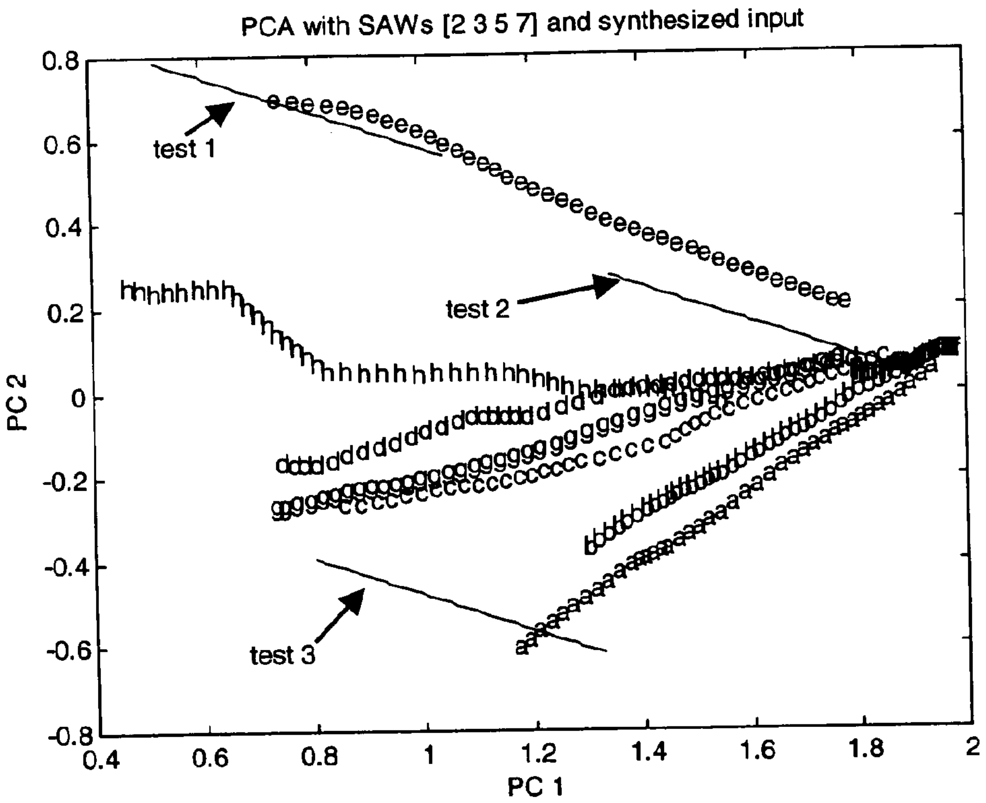

Figure 3. Test lines show the selectivity of the best-match approach used to synthesize replacement data for a faulted sensor. Each test line intersects with a specific analyte at a specific concentration. Apparent overlaps with other analytes are due to the dimensionality reduction of PCA. An ANN is able to uniquely identify the analytes.

Table 1. Confusion matrices showing neural network results. Actual inputs are listed in the leftmost column. ANN classification using the best-match approach to replacing faulted data is shown on the left, and the training data response is shown on the right for comparison. Results are averaged over 41 concentration steps. Zero values are suppressed for clarity.

	Test Data Outputs									Training Data Outputs							
	a	b	c	d	e	f	g	h		a	b	c	d	e	f	g	h
a	95.1	4.9								100							
b		100									100						
c			53.7			14.6		31.7				92.7			7.3		
d				63.4		26.8		9.8					100				
e					92.7	7.3								100			
f						100									100		
g								100								100	
h								100									100

The results for 10% noise are compared with the results for no added noise in Figure 4. Here we cluster the results for testing (no noise), training (no noise), testing (10% noise), and training (10% noise) in four concentration ranges. This figure shows the trend for all results to improve with increasing concentration, and it also shows the degradation caused by noise.

Table 2. CMDA values broken out by concentration range. There are 41 concentration values in the data set. Step 1 is the lowest concentration.

Concentration range	CMDA	CMDA combining *g* and *h*	CMDA of training data
1–10	56.8 %	64.9 %	96.6 %
11-20	71.6	81.8	100
21-30	87.5	100	100
31-41	87.5	100	100
1–41	75.6	86.4	99.1

Table 3. CMDA values broken out by concentration range at various added noise levels. There are 41 concentration values in the data set. Step 1 is the lowest concentration.

	testing data			training data		
Concentration range	1% noise	5% noise	10% noise	1% noise	5% noise	10% noise
1–10	67.8	55.6	46.5	93.6	78.6	63.0
11-20	80.6	70.8	57.4	97.1	87.0	72.6
21-30	99.8	88.9	69.9	100.0	91.6	81.0
31-41	100.0	94.8	79.4	100.0	91.3	80.0
1–41	87.3	77.9	74.4	97.7	87.2	51.3

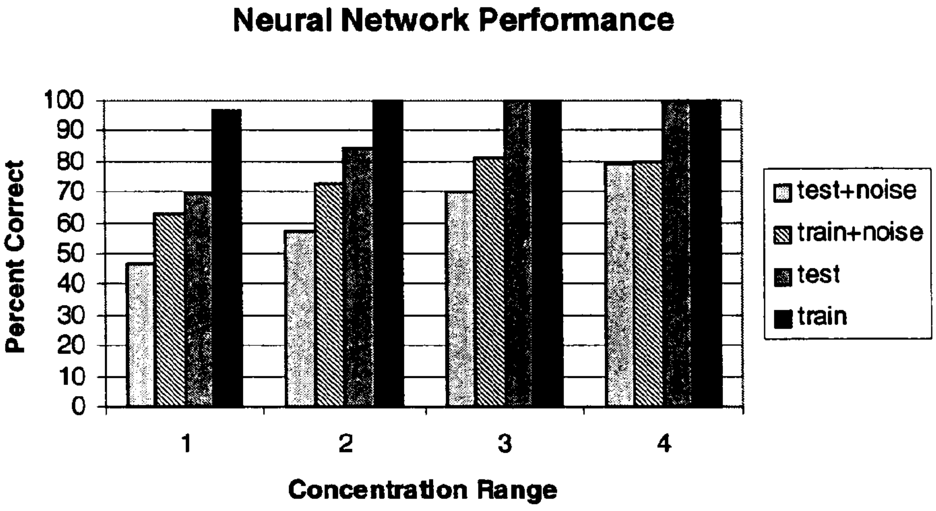

Figure 4. Percent correct identification by a neural network of eight analytes. The leftmost two bars in each cluster are the results for 10% added noise. Concentration range 1 is the lowest, 4 is the highest. Concentration steps are the same as shown in Table 3.

5. DISCUSSION AND CONCLUSIONS

The experimental results confirm the hypothesis that chemical identification by a sensor array can be improved by using an iterative "best-match" approach to replace sensor data that is known to be faulted. This technique is straightforward to implement, depending mainly on the ability to compute a confidence level for each neural network output in real time. The method of calculating this confidence level given here is quite simple, and could be readily implemented in hardware or software. This might be a useful technique to apply during a short transient period in which a standby sensor is recalibrated and switched in to replace a faulted main sensor, or during a longer period if there is no reason to expect the universe of measurement will drift significantly from the training data set.

REFERENCES

1. D. Wilson, T. Roppel, and R. Kalim, "Aggregation of sensory input for robust performance in chemical sensing microsystems," *Sensors and Actuators B* **64**, pp. 107–117, 2000.
2. T. Roppel, M. L. Padgett, S. Shaibani, and M. Kindell, "Robustness of a Neural Network Trained for Sensor Fault Detection, "*Proc. of the Workshop on Neural Networks,* 107-115, Auburn, Alabama, Feb. 11-13, 1991.
3. L. A. Klein, "A Boolean algebra approach to multiple sensor voting fusion," *IEEE Trans. Aero. and Elect. Sys.*29(2), pp. 317–327, April 1993.
4. P. Willett, M. Alford, and V. Vannicola, "The case for like-sensor predetection fusion," *IEEE Trans. Aero. and Elect. Sys.*30(4), pp. 986–1000, Oct. 1994.
5. E. Micheli–Tzanakou, *Supervised and Unsupervised Pattern Recognition*, CRC Press, ISBN 0-8493-2278-2, 2000
6. Aspen Technology Inc., Pittsburgh, PA.