CLUSTERING-BASED PATTERN RECOGNITION APPLIED TO CHEMICAL RECOGNITION USING SAW ARRAY SIGNALS

G. C. Osbourn, J. W. Bartholomew, G. C. Frye and A. J. Ricco

Sandia National Laboratories
Albuquerque, N.M. 87185

ABSTRACT

We present a new pattern recognition (PR) technique for chemical identification using arrays of microsensors. The technique relies on a new empirical approach to k-dimensional cluster analysis which incorporates measured human visual perceptions of difficult 2-dimensional clusters. The method can handle nonlinear SAW array data, detects both unexpected (outlier) and unreliable array responses, and has no user-adjustable parameters. We use this technique to guide the development of arrays of thin-film-coated SAW devices that produce optimal PR performance for distinguishing a variety of volatile organic compounds, organophosphonates and water.

INTRODUCTION

There is much current interest in the use of pattern recognition (PR) techniques to identify and quantify chemical analytes based on the multivariate responses of chemical microsensor arrays [1,2]. PR techniques used previously measured array responses of the class categories of interest, called training set data, to infer the class category associated with new array measurements (test data). The class categories can be the chemical identities of the analytes and can also represent the analyte concentrations. Each sensor in the array provides one dimension of the multidimensional data vectors. The application of PR techniques to microsensor-based chemical sensing is motivated by two advantages: (i) chemical identification is possible using only partially chemically selective microsensors rather than highly selective (and more difficult to develop) microsensors; (ii) PR can in principle identify a large number of chemical species using a small, fixed set of sensors.

The desired arrays of sensors are those that yield distinctive response patterns for different chemicals, i.e. so that the clusters of pattern vectors associated with different chemical classes are spatially separated and distinguishable in the pattern vector space. However, not all arrays yield useful separations of the chemical classes, so much of the work required to develop a useful chemical pattern recognition scheme is in finding a useful combination of chemical sensors.

Surface acoustic wave (SAW) devices that have been coated with different partially selective films are of current interest for array applications [2-4]. The variety of coatings that can be fabricated suggests the possibility of detecting a wide range of analytes with SAW devices. However, coated SAW devices often respond nonlinearly and occasionally nonmonotonically as a function of analyte concentration [5]. These devices also do not necessarily yield additive responses to chemical mixtures. Such responses present severe difficulties for conventional chemometric techniques, which work well for linear response signals.

We describe a new PR technique which can treat nonlinear and nonmonotonic sensor responses. The technique is computationally efficient, provides warning for new data points that are atypical or that cannot be reliably identified, and requires no user-adjustable parameters. We demonstrate the technique with the optimization of arrays of coated SAW devices for identifying volatile organic compounds (VOCs), organophosphonates and water. VOC detection and monitoring is of interest for a number of industrial and DOE applications [4] and the organophosphonates are simulants for chemical warfare agents [3]. We describe the selection of SAW coating combinations and compute estimates of the expected performance of the arrays on new data. The arrays are developed and demonstrated using chemicals presented individually. The performance of this approach for array responses to complex mixtures remains the subject of future study.

PATTERN RECOGNITION TECHNIQUE

Our technique relies on a new empirical approach to k-dimensional cluster analysis that incorporates measured human visual perceptions of difficult 2-dimensional clusters [6]. We use this new clustering method because it: (i) mimics human cluster perception; (ii) requires no prior knowledge about the final cluster result (e.g. the number of clusters); (iii) outperforms commercially available cluster methods on benchmarking tests [6]. Here we describe the use of the clustering results to carry out PR or to evaluate the expected PR performance using sets of coated SAW sensor responses. Details of the clustering method are described in Ref. 6.

PR using the cluster results is illustrated in Fig. 1a. The open and closed circles represent hypothetical training data for two classes and the × symbols represent hypothetical test data points of unknown chemical class identity. The two dimensions of the data correspond to simultaneous responses of two different sensors. Computed clusterings of the test points are indicated by the lines connecting the points. Three possible cases occur for the class assignment of each new test point: (i) The new point is clustered only with training data of a single class and is assigned that same class -- this is the desired result; (ii) The new point is clustered with none of the training data and is not assigned a class -- these outlier points can result from detection of a new chemical class that is not present in the training data set, although sensor measurement errors can also produce outliers; (iii) The new point is clustered with training points from multiple classes -- these points occur where the training set classes overlap and cannot be reliably distinguished, so that the class assignment is ambiguous. We note that cases (ii) and (iii) provide useful information that is often unavailable or unreliable from existing PR techniques.

It is usually necessary to estimate the expected performance of the sensor array using only the measured training data set. We use the "leave-one-out" approach [7], where each data point is individually removed from the complete training set and examined as a new test point by the PR technique. This approach maximizes the use of the available training data and avoids the favorable bias of the results that occurs by including the point that is being classified. This approach is not always convenient in some PR methods that require additional computations for each version of the training set with a point removed (e.g. retraining a neural net). Fig. 1b illustrates how the cluster results are used for the leave-one-out computations. The known class identity of each training data point is compared with the result inferred from the clustering approach shown in Fig. 1a. There are again three possible cases which can arise: (i) The point is clustered only with other training data of the correct class -- such points are correctly identified; (ii) The point is clustered with none of the other training data -- these outlier points are too far away from other points in the same class for clustering to occur. Outlier points in the leave-one-out analysis are usually due to inadequate sampling of the class in the parts of pattern vector space near the outliers, and suggest the need for more training points. However, these can also result from sensor measurement errors; (iii) The point is clustered with other training points from one or more classes that do not match the correct class. These points identify overlaps between classes in pattern vector space where the sensor array is unable to reliably

0-9640024-0-X/hh1994/$20©1994TRF
DOI 10.31438/trf.hh1994.44

Solid-State Sensors, Actuators, and Microsystems Workshop
Hilton Head Island, South Carolina, June 12-16, 1994

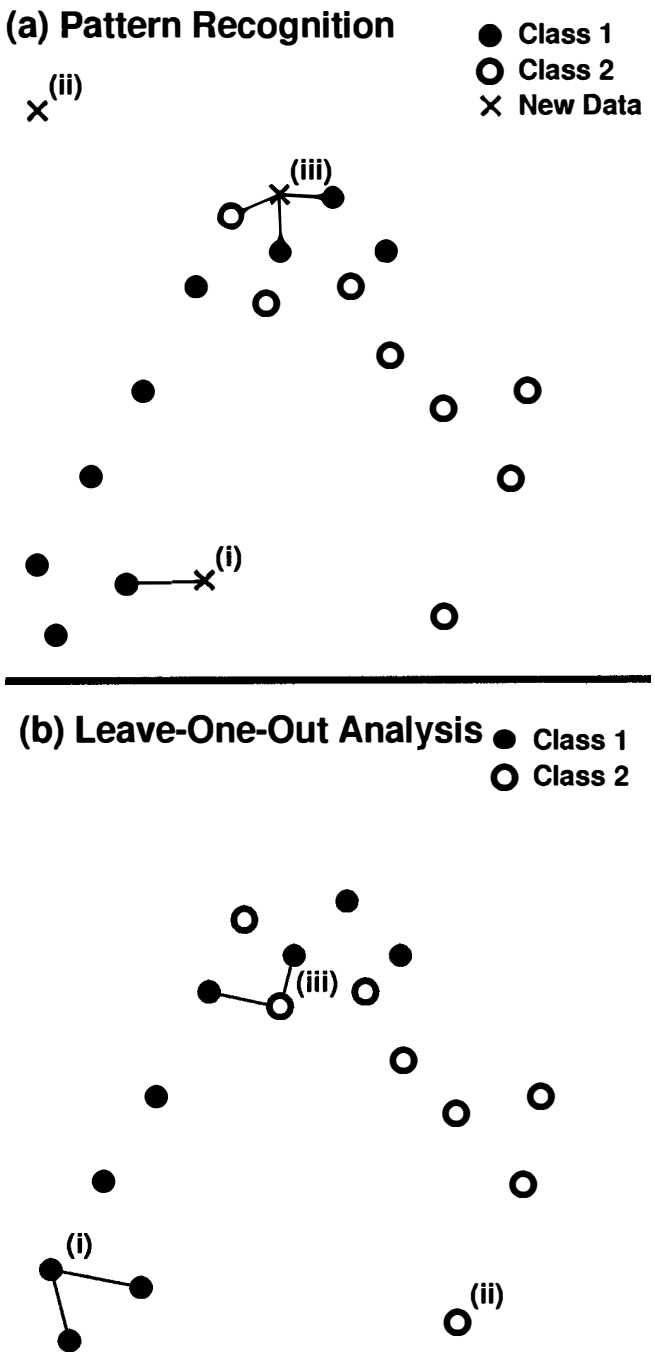

Fig 1. (a) Illustration of the use of cluster results for PR. Open and closed circles are training data, × symbols are test data. Computed clusterings of test points with training data are indicated by connecting lines. The three cases illustrated are: (i) recognized; (ii) outlier; and (iii) ambiguous or incorrect. (b) Use of cluster results for leave-one-out PR analysis on the same training data. The three illustrated cases are: (i) correctly recognized; (ii) outlier; and (iii) incorrect.

distinguish chemical identities. Statistics on these three cases for the entire training set provide three separate figures of merit for the sensor array under consideration, i.e. the percentage correctly identified, the percentage that indicates inadequate training set sampling, and the percentage that indicates class overlap. These figures of merit are used here to optimize the selection of the array sensors.

EXPERIMENTAL PROCEDURES

We analyze experimental results from two sets of coated-SAW arrays and analytes. In both studies we make use of only the frequency shifts which result from the uptake of the analyte by the SAW device coatings. Other measurable responses, e.g. attenuation [4], are also available for PR, and we will examine these in future work.

The first set of chemicals includes water, an alcohol (isopropanol), a ketone (acetone), non-chlorinated hydrocarbons (d-limonene, n-hexane, and toluene) and chlorinated hydrocarbons (CCl_4, TCE, and chloroform). These analytes are sensed using SAW devices coated with ethyl cellulose, polyisobutylene and nafion polymer films. The data are acquired by alternating between a chemical-free purge stream and a stream at a constant chemical concentration. Chemical concentration is controlled by the relative flow rates of a chemical-free mix-down stream and a saturated stream (obtained by passage through a bubbler containing the chemical). Four concentrations per chemical were obtained. This produced 12 data points for both the chlorinated and non-chlorinated classes.

The second set of chemicals includes water, organophosphonates (DIMP, DMMP), a ketone (acetone), aromatic hydrocarbons (benzene, toluene), chlorinated hydrocarbons (CCl_4, TCE), aliphatic hydrocarbons (cyclohexane, isooctane), and alcohols (methanol, pinacolyl alcohol, n-proponal). These analytes are sensed using SAW devices coated with nine self-assembled monolayer (SAM) films and plasma-grafted polymer films (PGFs). These films and the data acquisition method are described in a companion paper [8]. We retained 21 points per chemical, with concentrations ranging from 9% to 49% of the saturation vapor pressure.

DATA PREPROCESSING FOR CLUSTERING-BASED PATTERN RECOGNITION

It is common for coated SAW devices to yield ranges of responses that are quite different in magnitude. In order for each device to have a comparable effect on the class separations in pattern vector space, the magnitudes of the device responses must be comparable. We equalize the responses of each sensor by a multiplicative scale factor, so that the largest absolute training data response of each sensor is unity.

The coated SAW sensor responses and the lengths of the multivariate pattern vectors generally increase as the chemical concentrations are increased. This tends to spread the class vectors away from the origin and each other as the chemical concentrations are increased. We improve the clusterings of the sensor data by normalizing the pattern vectors to unit length after the device responses have been equalized. A comparison of PR performance for the raw and normalized data are described below.

DETERMINATION OF OPTIMAL ARRAYS

A key element in the successful development of a pattern recognition system is the selection of a useful set of measurements that can separate the classes in pattern vector space. This corresponds to selecting a set of SAW coatings and a set of chemical concentrations to include in the training set. One might expect that arrays with larger numbers of distinct sensor signals, k, would always be preferable for PR applications. This is not true for two reasons. From a practical viewpoint, arrays with the fewest devices are likely to require simpler fabrication and data acquisition procedures. This motivates the use of the smallest arrays that provide the desired performance. From a PR performance viewpoint, discriminating ability peaks and then declines as k is increased beyond approximately N/2, where N is the number of training data points per class. The reasons for this well-known behavior of PR systems are discussed in Ref. 7. Here we emphasize that the number of SAW films that can be usefully included in an array is limited by the number of chemical concentrations per class that is included in the training data set. This well-known condition for reliable PR analysis of arbitrary k-dimensional data is often not satisfied in published sensor PR work. We also note that estimates of PR performance that

do not remove examined points from the training data when computing their class identities, i.e. that do not use the leave-one-out approach, will exhibit increasingly perfect performance as k increases. Such biased estimates are not representative of the PR performance on future data.

For large sets of distinct devices, we must search for the smallest subsets that produce good separation of the chemical classes of interest in a leave-one-out analysis. We do this directly for the nine SAMs and PGFs by examining all combinatorial sets of arrays. The computations are done on a SPARC 10 workstation. The primary figure of merit we seek to minimize is the percentage of incorrect leave-one-out points in the data set. We also seek to minimize the number of outliers among the array combinations with the smallest incorrect percentages. The number of data points per chemical class is sufficient to usefully include all of the SAM-coated and PGF-coated SAW devices in the array, but we are interested in identifying the smallest subset of devices with the best PR performance.

RESULTS

The leave-one-out results for the ethyl cellulose, polyisobutylene and nafion coatings yield no incorrect points for the class combinations chosen. The locations of the training points are shown in the plot of Fig. 2. Note that this is one perspective of a

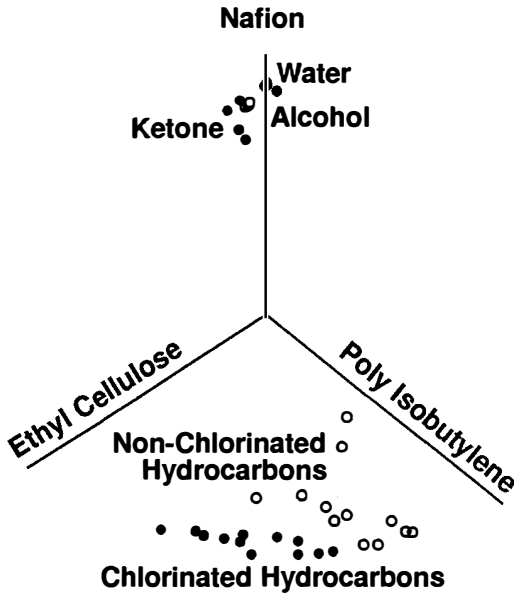

Fig 2. One view of a three-dimensional pattern vector space plot of 3-film SAW array data. All of these chemical classes are distinguished. The alcohol and ketone classes are seen to separate when viewed along the nafion axis. The 3 types of chemicals within the chlorinated and non-chlorinated classes are too similar to be distinguished with this array.

three-dimensional plot. These results show that a variety of chemical classes can be distinguished using these three SAW coatings. The ketone and alcohol classes appear to overlap from this viewing orientation, but can be seen as distinct when viewed along the nafion axis direction. The three individual chemicals within the chlorinated and non-chlorinated VOC classes are not reliably separated, so that additional measurements are needed to distinguish these chemicals. However, more concentration values are required for the individual chemical classes to achieve this improvement.

Figs. 3a-d show the statistics obtained from the nine SAMs and PGFs. The number of array combinations that yields a particular percentage of incorrect identifications is plotted for arrays with 3 to 6 films in Figs. 3a-d, respectively. Again it is seen that all three-film

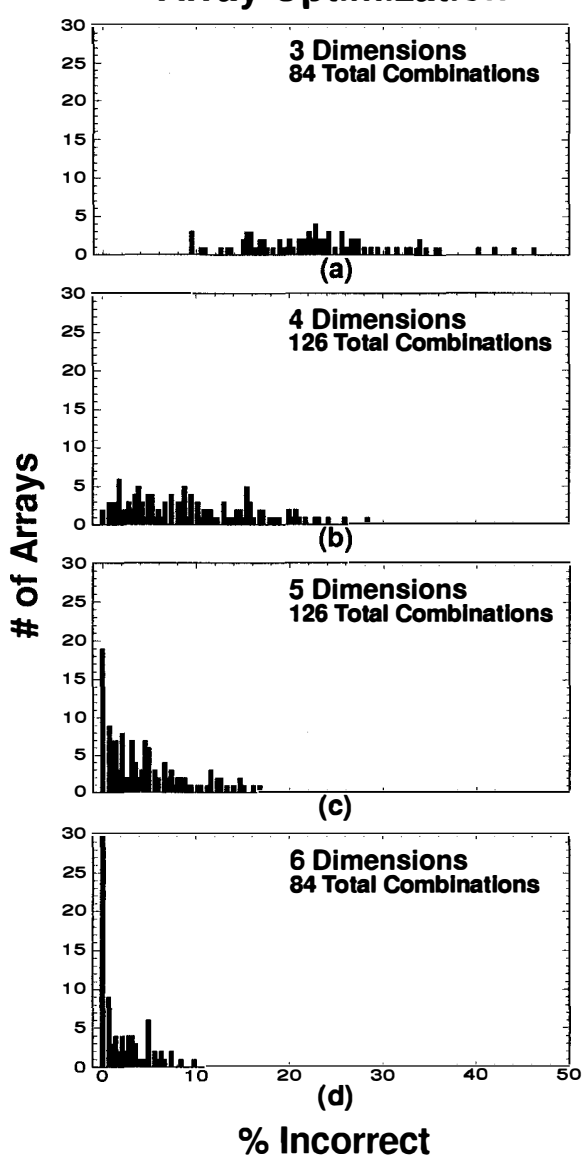

Fig 3. Histograms of number of arrays vs percentage incorrect from the leave-one-out analyses for the SAM-coated and PGF-coated SAWs. Panels (a)-(d) show results for arrays with three to six films per array, respectively.

arrays generate some incorrect chemical identifications. The overall PR performance improves as films are added, and the number of array combinations exhibiting no incorrect responses is nonzero for four or more films in the arrays. The best three-to-six film combinations are [$HS(CH_2)_{10}COO^-/Cu^{2+}$, Eugenol-30, PIB-PGAA 5+15], [$HS(CH_2)_{15}CH_3$, PGAA-PGAA 5+15, Eugenol-30, PIB-PGAA 1+15], [$HS(CH_2)_{15}CH_3$, $HS(CH_2)_{15}CH_3$, PGAA-PGAA 5+15, PGAA-PGAA 5+30, PIB-PGAA 1+15], [$HS(CH_2)_{10}COOH$, $HS(CH_2)_{10}COO^-/Cu^{2+}$, $HS(CH_2)_{15}CH_3$, Eugenol-30, PIB-PGAA 1+15, PIB-PGAA 5+15], respectively (see ref. 8 for an explanation of the film abbreviations). Since the higher-dimensional results can not be graphically displayed, we show the best three-film case in Fig. 4. This three-film set is not able to separate all of the chemicals. For clarity, we show only some of the chemical classes which are non-overlapping in Fig. 4.

Finally, we have examined the effect of normalizing the pattern vector lengths on these results. This is done by computing results with

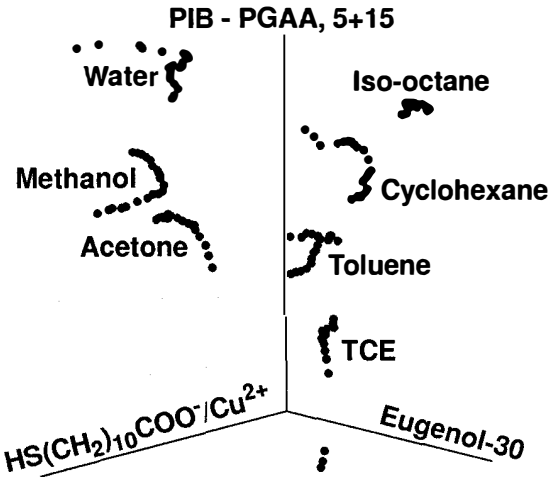

Fig 4. One view of a three-dimensional pattern vector space plot of the best 3-film SAW array from Fig. 3a. Not all of the individual chemicals are correctly identified, and for clarity we show only some of the non-overlapping chemical data.

a range of scaling values ranging from 1, corresponding to raw data, to 1/(vector length), corresponding to a completely normalized vector length. Fig. 5 plots the number of incorrect responses as a function of the degree of normalization for the three-film array shown in Fig. 4. The PR performance is best for nearly normalized vector lengths and is worst for the raw data.

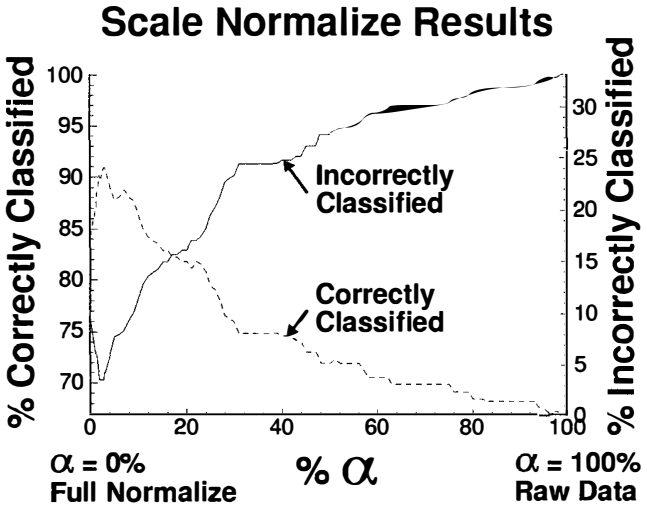

Fig 5. Effect of the amount of pattern vector normalization on the results of the three-film array of Fig. 4. Each pattern vector is multiplied by 1/(L+α(1-L)), where L is the length of the pattern vector and α determines the degree of normalization.

CONCLUSIONS

We present a new PR technique for chemical identification using arrays of microsensors. The technique relies on a new empirical approach to k-dimensional cluster analysis which incorporates measured human visual perceptions of difficult 2-dimensional clusters. The method can handle nonlinear SAW array data, detects both unexpected (outlier) and unreliable array responses, and has no user-adjustable parameters. We use this technique to develop small arrays of thin-film-coated SAW devices which produce optimal PR performance for distinguishing a variety of volatile organic compounds, organophosphonates and water. We find that three-film arrays are sufficient to distinguish a number of distinct chemical classes, but are unable to distinguish closely similar chemicals in the non-chlorinated and chlorinated hydrocarbon classes. For vapor concentrations greater than 9% of saturation, arrays with as few as 4 selected films give 100% correct classification for these chemical classes and the individual chemicals. In addition, many different five-film combinations give 90-100% correct identifications for the individual chemicals.

ACKNOWLEDGEMENT

Work supported by the US DOE under contract DE-AC04-94AL85000.

REFERENCES

(1) X. Wang, J. Fang, P. Carey, S. Yee, "Mixture analysis of organic solvents using nonselective and nonlinear Taguchi gas sensors with artificial neural networks", Sensors and Actuators B, 13, 455 (1993).
(2) S. L. Rose-Pehrsson et. al., "Detection of hazardous vapors including mixtures using pattern recognition analysis of responses from surface acoustic wave devices", Anal. Chem 60, 2801 (1988).
(3) A. J. Ricco, C. Xu, R. M. Crooks, and R. E. Allred, "Chemically sensitive interfaces on SAW devices" in *Chemically Sensitive Interfaces*, ACS Symposium Series No. 561, Amer. Chem. Society, Washington, D. C. (1994), in press.
(4) G. C. Frye, S. J. Martin R. W. Chernosek and K. B. Pfeifer, "Portable acoustic wave sensor systems for on-line monitoring of volatile organics", Int. J. of Environ. Conscious Manufacturing, 1, 37 (1992).
(5) G. C. Frye and S. J. Martin, "Velocity and attenuation effects in acoustic wave chemical sensors", Proc. 1193 Ultrasonics Symposium (IEEE, 1993) p.379.
(6) G. C. Osbourn and R. F. Martinez, "Empirically defined regions of influence for clustering analyses", Sandia Report SAND 94-0210.
(7) D. J. Hand, *Discrimination and Classification*, John Wiley and Sons, New York 1981.
(8) A. J. Ricco et. al., this proceedings.