Semiconductor Physics, Quantum Electronics & Optoelectronics, 2010. V. 13, N 4. P. 393-397.

PACS 07.07.Df

Methods of cluster analysis in sensor engineering: advantages and faults

Yu.V. Burlachenko, B.A. Snopok
V. Lashkaryov Institute of Semiconductor Physics, NAS Ukraine
41 Prospect Nauky, Kyiv 03028, Ukraine
Tel.: (*380-44*) *525-52-46; e-mail: b_snopok@yahoo.com*

Abstract. We consider the crisp and fuzzy partitioning techniques of cluster analysis bearing in mind their application for classification of data obtained with chemical sensor arrays. The advantage of the cluster analysis techniques is existence of a parameter $S(i)$. This parameter gives quantitative efficiency of classification and can be used as optimization criterion for sensor system as a whole as well as the measurement procedure. The crisp and fuzzy techniques give practically the same result when analyzing the data that cluster uniquely. It is shown that big value of the parameter $S(i)$ is not sufficient for adequate data partitioning into cluster in more complicated cases, and the results of clusterization for the above techniques may diverge. In this case, one should apply both techniques concurrently, checking the correctness of partitioning into clusters against the principal component analysis.

Keywords: multisensor systems, cluster methods, recognition, classification.

Manuscript received 12.01.10; revised manuscript received 01.02.10; accepted for publication 02.12.10; published online 30.12.10.

1. Introduction

The development and fabrication of an electronic analog to biological nose is one of the most interesting practical tasks of modern science. In recent years, much progress has been achieved in this area. Now there are many developments and commercial devices called Electronic Noses (EN) [1]. However, contrary to biological nose that provides an organism with all necessary information on the character of odors from its nearest neighborhood, EN gives partial information only. Indeed, the multisensor arrays that serve as basis for EN have a certain selectivity profile. Therefore, each device of that type can be applied to solve only limited range of tasks.

Choice of the most efficient sensor array for solution of a specific problem is one of the most important tasks in optimization of EN-type devices. Various generalized mathematical models [2-4], statistical approaches [5, 6], estimations of information content via the Fisher information [7, 8], predictions based on system dynamic behavior [9] etc. have been proposed in this area and successfully demonstrated in many cases [9]. All the above approaches are based, up to a point, on statistical calculations most of which use the cluster analysis techniques. When classifying data, various versions of pattern recognition procedures are used to predict the properties of an object that were not measured directly (chemical composition) but are related indirectly to measurements via unknown or undetermined interrelations.

To estimate the device operation efficiency when solving a task, one should define a criterion for such estimation. The objective of any EN-type sensor system is classification with further recognition of the objects studied (generally, these are multicomponent mixtures). Strictly speaking, just classification efficiency could serve as criterion of array optimization. Moreover, having a quantitative estimate of classification efficiency, one could optimize not only the array itself but the measurement procedure as well, thus ensuring the choice of the most informative part of response.

However, only some techniques of cluster analysis make it possible to estimate classification efficiency *quantitatively*. In this work, we consider the partitioning techniques of cluster analysis from the viewpoint of their application for classification of data from multisensor arrays. The advantage of these techniques is existence of a parameter $S(i)$ that expresses the classification efficiency quantitatively. We consider appropriateness of this parameter as criterion of sensor array optimization. The peculiarities of application of partitioning methods in sensor technique are also considered.

2. The partitioning methods of cluster analysis

Classification means partitioning of a set of objects or observations into uniform groups (clusters) whose elements are similar, while there are quantitative distinctions between elements belonging to different clusters [10]. Thus, the objective of cluster analysis is structuring the multidimensional input data and attribution of every object from the given set to one of the clusters. The classical methods of cluster analysis (crisp techniques) lead to partitioning of the data set into clusters with well-defined boundaries. This means that, whatever the input data, they should be ascribed to a certain class. Contrary to the crisp techniques, those based on the concept of "fuzzy logic" calculate a membership for each object, which indicates how strongly the object belongs to a cluster. Thus, assignment of an object to a certain class is presented as something true up to a point only.

Different techniques of cluster analysis are integrated into most of the modern software packages used for statistical data processing. This makes application of such packages simple and obvious. In this work, comparison of the results of data classification for multisensor arrays is made using the S-PLUS software environment, with the partition around medoids (PAM) and cluster analysis in the fuzzy logic format serving as examples.

Let us start consideration with PAM. This technique belongs to the crisp methods: each object is assigned to one cluster only. The technique is based on search for a certain number of representative objects called medoids. The latter are chosen in such a way that the dissimilarities between all objects and their nearest medoid are minimal. The number of clusters is set by the user. S-PLUS has an option of visualization of the results of objects partitioning into clusters through construction of a cluster plot (clusplot).

For each i-th objects, a parameter $s(i)$ is calculated, which characterizes quality of that object clusterization. Let us dwell on the physical sense of that parameter, without going into details [11]. The value of $s(i)$ may be interpreted in the following way:

- $s(i) \approx 1$ – the i-th object is classified well (into the given cluster);
- $s(i) \approx 0$ – the i-th object is between two clusters;
- $s(i) \approx -1$ – the i-th object is classified badly (belongs to another cluster rather than the given one).

The $s(i)$ values for all objects are plotted in a special diagram (the so-called silhouette plot). In this case, all the objects are partitioned into groups, depending on their assignment to a certain cluster. The average value S for all clusters in the silhouette plot is a parameter that characterizes quantitatively the classification quality as a whole (for all objects). It is this parameter that may be applied for optimization of sensor arrays.

Now let us consider the fuzzy partitioning technique that is based on the concept of "fuzzy logic". Contrary to PAM where assignment of an object to a certain class is either 0 or 1, in the fuzzy partitioning technique it may take any value from 0 up to 1. The results of analysis with the fuzzy partitioning technique also may be presented as a clusplot and silhouette plot; to this end, the closest crisp partitioning is chosen. As a rule, the results obtained with fuzzy partitioning and PAM are the same, if separation of the objects into classes is sufficiently unambiguous. If, however, there are some objects in a data array whose assignment to a certain class is not well-defined, then different techniques may give different results. Therefore, it seems to be of importance that comparative analysis of adequacy of these data classification should be made with the crisp as well as fuzzy techniques.

3. Experimental

A set of experimental data was obtained using an array of three QCM sensors (AT-cut quartz resonators with resonance frequency of 10 MHz) modified with phthalocyanine (H_2Pc, CuPc, PbPc) films 100 nm thick. The following analytes were used: (1) ethanol; (2) triethylamine; (3) propylamine; and (4) water. Three repeated measurements were performed with each analyte. For the features of the measurement procedure as well as the experimental set-up design see [12].

4. An example of application of PAM and fuzzy partitioning

Figure 1 shows, as an example, the typical experimental curves for three sensors exposed to ethanol vapor (these curves were used in further calculations). Table 1 presents the normalized values of sensor responses, S_{nm}, at a moment t = 35 s since the beginning of measurements:

$$S_{nm} = \frac{F_{nm}}{\sum_n F_{nm}} .$$

Here, F_{nm} is the response of the n-th sensor (taken in the m-th measurement) to the same analyte; m = 1…3 numbers of measurements, while n = 1…3 numbers of sensors.

As was noted earlier, availability of a priori information on the number of classes is presumed when applying the cluster methods. At the same time, in many cases it is necessary to evaluate data quality bearing in mind possible classification (how the data are clasterized per se). To solve this task, one usually applies the principal component analysis (PCA) [10]. This makes it possible to project the response space onto a plane with minimum distortions, visualize the data in the transformed space of sensor coordinates, and estimate qualitatively the degree of inherent data clusterization. The data from Table 1 obtained with PCA are presented in Fig. 2. (The numbers correspond to those of experiments.)

Table 1. Normalized sensor responses S_{nm} at t = 35 s since the beginning of measurement.

Analyte	Measurement #	Sensor 1	Sensor 2	Sensor 3
Ethanol	1	0.316058	0.270269	0.413674
	2	0.274731	0.273789	0.451479
	3	0.311210	0.280612	0.408178
Triethylamine	4	0.133913	0.564111	0.301976
	5	0.154681	0.564090	0.281228
	6	0.149661	0.553314	0.297026
Propylamine	7	0.322383	0.205610	0.472007
	8	0.294352	0.211957	0.493691
	9	0.298338	0.202595	0.499067

One can see from Fig. 2 that the objects (4, 5, 6) – triethylamine – make a clearly pronounced separate group. The object 2 (ethanol) is closer rather to the group (7, 8, 9) than its own class (1, 3). This fact makes the task of its correct classification much more difficult. We used intentionally the data whose classification is not apparent. Our aim was to demonstrate the fact that, in such a situation, two different cluster analysis techniques may give different results.

Shown in Fig. 3 are the silhouette plot and clusplot constructed with PAM using the data from Table 1. One can see that the object 2 (ethanol) is assigned to the class (7, 8, 9) - propylamine, just as according to PCA, i.e., its classification is wrong. In fact, negative $s(i)$ value for this object suggests that it belongs to another class rather than this one.

Figure 4 presents a silhouette plot constructed with the fuzzy partitioning technique using the same data. In this case, all the objects are in their own classes. One can see from Fig. 4 that just the second technique gives correct classification. This is in spite of the fact that S takes a bigger value in the first case rather than the second one (0.67 for PAM and 0.66 for fuzzy partitioning). Thus, one can state that bigger S value is a necessary but not sufficient condition for correct data classification.

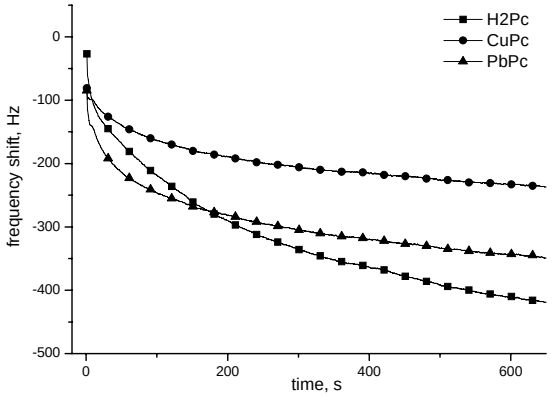

Fig. 1. Responses to ethanol vapor of QCM-sensors coated with 100 nm films of phthalocyanines (H_2Pc, CuPc and PbPc).

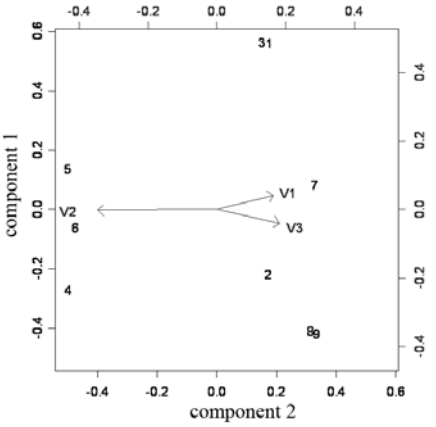

Fig. 2. PCA plot related to the responses of a three-sensor array to ethanol, triethylamine and propylamine.

It should be noted that analysis of different situations with data classification for multisensor arrays testifies unambiguously that one cannot say in advance what technique (crisp or fuzzy) will be more appropriate for consideration of a specific case. Therefore, it seems reasonable to perform classification using both techniques in parallel to improve reliability of results. In this case, the PCA method can serve for both visualization and check of the results given by the cluster analysis techniques because, contrary to the partitioning techniques, it does not require availability of a priori information on the number of clusters.

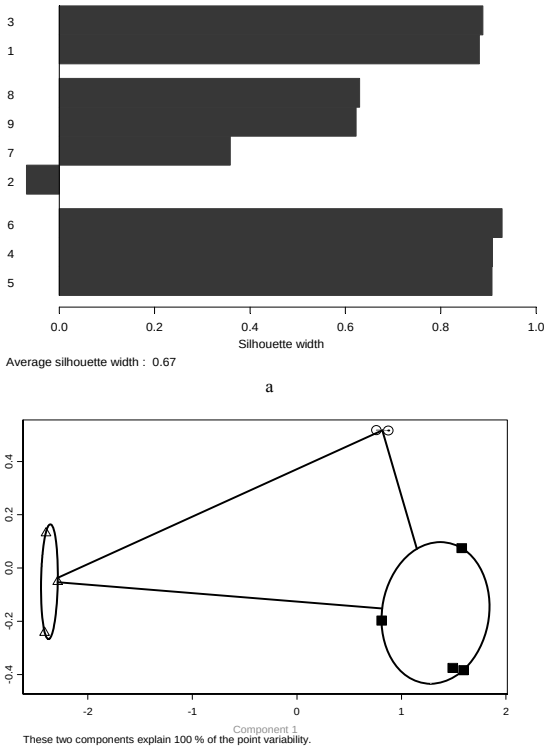

Fig. 3. Silhouette plot (a) and clusplot (b) constructed according to PAM using sensor responses to ethanol, triethylamine and propylamine.

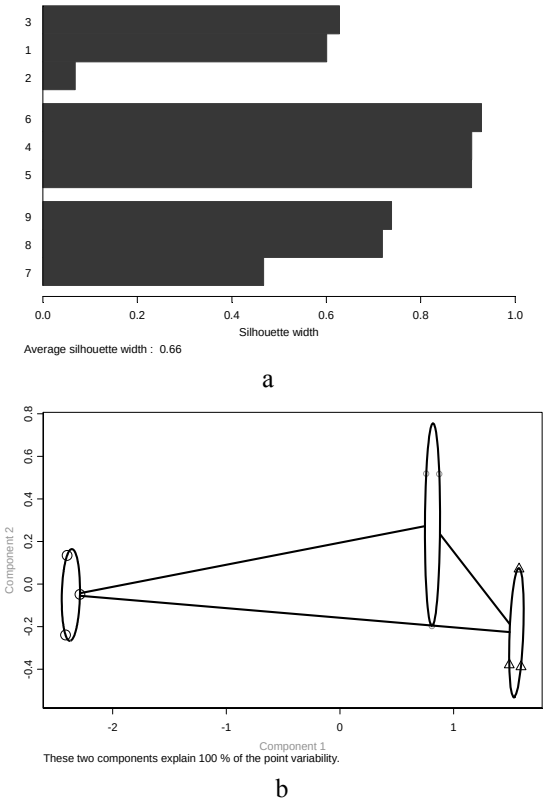

Fig. 4. As in Fig. 3 but made with the fuzzy partitioning technique.

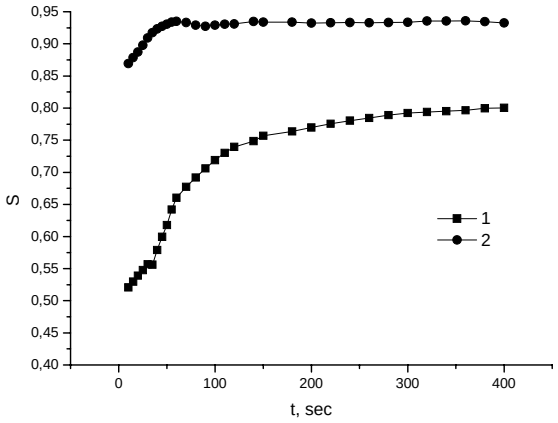

Fig. 5. The classification efficiency curves $S(t)$ constructed using the sensor array responses to analytes: 1 - ethanol, triethylamine and propylamine; 2 - water, triethylamine and propylamine.

5. Use of parameter *S* for optimization of sensor array and measurement procedure

The parameter S can serve not only for array optimization (i.e., for comparison of efficiencies of individual sensors in an array) but for choice of the most representative (i.e., ensuring the best classification) region of response surface as well. Classification efficiency may vary considerably in the course of time of measurement. The reasons for this are the effects of kinetic discrimination, on the one hand, and those of reproducibility (different portions of adsorption curve are affected differently by the external factors), on the other hand [13]. Indeed, the calculation of S value for every instant of time of experiment makes it possible to obtain time dependence of classification efficiency, $S(t)$.

Shown in Fig. 5 are the examples of such dependences for classification of two sets of analytes: ethanol, triethylamine and propylamine (curve 1) and water, triethylamine and propylamine (curve 2). The same multisensor arrays were used in both cases. Such a presentation makes it possible to determine the most informative (from the viewpoint of analyte distinctive features) part of array response with respect to any of the analytes used. To illustrate, for the first set of analytes, it seems more reasonable to consider the stationary response amplitudes (the peak of $S(t)$ dependence is in the saturation region of the adsorption curves). At the same time, for the second set of analytes (that differed from the first one by a single analyte only), the peak of discrimination efficiency is observed in the kinetic region. (Note once more that the same multisensor array was used in both cases.)

Of course, the curve $S(t)$ can be used only after check for classification adequacy in different points of the curve. To this end, one should construct silhouette plots for sampling instants using the partitioning techniques. It is expedient to recall here that, as shown earlier, PAM gives wrong classification for the first set of analytes at $t = 35$ s (see Fig. 3).

6. Conclusions

The approaches of mathematical statistics and experiment optimization are widely used in analytical chemistry practice to obtain information from big analytical data arrays. The techniques of cluster analysis are necessary and extremely convenient tool for solving such tasks with respect to multidimensional data obtained with sensor arrays made for various purposes. As a rule, the results of classification obtained with the crisp and fuzzy techniques coincide. If, however, the data are classified ambiguously, it is reasonable to apply both approaches in parallel, checking the result obtained with PCA. In this case, availability of the parameter S makes it possible to use the cluster methods for sensor array optimization as well as choice of the most informative region of the response surface. This will enable one to increase efficiency of the analytical procedures based on multisensor arrays to solve various tasks of gas analysis via minimization of costs and time required for analytical signal measurement and extracting chemical information on analyte using the databases for reference specimens.

Acknowledgements

This work got a financial support from the National Academy of Sciences of Ukraine.

References

1. M. Peris, L. Escuder-Gilabert, A 21st century technique for food control: Electronic noses // *Analytica Chimica Acta* **638**(1), p. 1-15 (2009).
2. P.W. Carey, B.R. Kowalski, Chemical piezoelectric sensor and sensor array characterisation // *Analytical chemistry* **58**, p.3077-84 (1986).
3. P.W. Carey, K.R. Beebe, B.R. Kowalski, Selection of adsorbates for chemical sensor arrays by pattern recognition // *Analytical chemistry* **58**, p.149-53 (1986).
4. P.W. Carey, K.R. Beebe, B.R. Kowalski, Multicomponent analysis using an array of piezoelectric crystal sensors // *Analytical chemistry* **59**, p.1529-34 (1987).
5. S.M. Briglin, M.S. Freund, P. Tokumaru, N.S. Lewis, Exploitation of spatiotemporal information and geometric optimization of signal/noise performance using arrays of carbon black-polymer composite vapor detectors // *Sensors and Actuators B: Chemical* **82**(1), p. 54-74 (2002).
6. K.J. Albert, N.S. Lewis, C.L. Schauer, G.A. Sotzing, S.E. Stitzel, T.P. Vaid, D.R. Walt, Cross-reactive chemical sensor arrays // *Chem Rev.* **100**(7), p. 2595-626 (2000).
7. M.A. Sànchez-Montañès, T.Pearce, Fisher information and optimal odor sensors // *Neurocomputing* **38-40**, p. 335-341 (2001).
8. T.C. Pearce, P.F.M.J. Verschure, J. White, J.S. Kauer, Stimulus encoding during the early stages of olfactory processing: A modeling study using an artificial olfactory system // *Neurocomputing* **38**, p. 299-306 (2001).
9. B.A. Snopok, I.V. Kruglenko, Nonexponential relaxations in sensor arrays: forecasting strategy for electronic nose performance // *Sensors and Actuators B: Chemical* **106**(1), p. 101-113 **(**2005).
10. P.C. Jurs, G.A. Bakken, H.E. McClelland, Computational methods for the analysis of chemical sensor array data from volatile analytes // *Chem. Rev.* **100**, p. 2649-2678 (2000).
11. A. Struyf, M. Hubert, P.J. Rousseeuw, Integrating robust clustering techniques in S-PLUS // *Computational Statistics and Data Analysis* **26**, p. 17-37 (1997**)**.
12. Yu.V. Burlachenko, B.A. Snopok, Multisensor arrays for gas analysis based on photosensitive organic materials: An increase in the discriminating capacity under selective illumination conditions // *Journal of Analytical Chemistry* **63**(6), p. 557-565 **(**2008).
13. B.A. Snopok, I.V. Kruglenko, Multisensor systems for chemical analysis: state-of-the-art in Electronic Nose technology and new trends in machine olfaction // *Thin Solid Films* **418**(1), p. 21-41 (2002**).**