СЕНСОРИ ТА ІНФОРМАЦІЙНІ СИСТЕМИ

SENSORS AND INFORMATION SYSTEMS

УДК 681.545

APPLICATION OF LINEAR CLASSIFIERS FOR ANALYSIS OF THE SENSOR ARRAYS DESCRIPTIVENESS FOR DETECTION OF THE VOLATILE COMPOUNDS MOLECULES

A. S. Pavluchenko, Z. I. Kazantseva, I. A. Koshets, Yu. M. Shirshov

V. Lashkaryov Institute of Semiconductor Physics of NASU
Tel. /fax: (+380-44) 2651827, (+380-44) 2655626
e-mail: shirshov@isp. kiev. ua

Summary

APPLICATION OF LINEAR CLASSIFIERS FOR ANALYSIS OF THE SENSOR ARRAYS DESCRIPTIVENESS FOR DETECTION OF THE VOLATILE COMPOUNDS MOLECULES

A. S. Pavluchenko, Z. I. Kazantseva, I. A. Koshets, Yu. M. Shirshov.

In this work some approaches to improvement of the chemical images classification performance due to preprocessing and dimensionality reduction of the multisensor array responses are studied. A criterion for linear separability of the analyte classes is proposed, allowing to choose an optimal method of responses processing for the particular classification task without the need of cross-validation involving multiple classifier retraining.

Keywords: multisensor array, quartz crystal microbalance, linear classifier, feature space reduction, separability criterion.

Анотація

ЗАСТОСУВАННЯ ЛІНІЙНИХ КЛАСИФІКАТОРІВ ДЛЯ АНАЛІЗУ ІНФОРМАТИВНОСТІ СЕНСОРНИХ МАСИВІВ ПРИ РЕЄСТРАЦІЇ МОЛЕКУЛ ЛЕТКИХ РЕЧОВИН

Павлюченко О. С., Казанцева З. І., Кошець І. А., Ширшов Ю. М.

В роботі розглянуті деякі підходи до покращення якості класифікації при розпізнаванні хімічних образів за рахунок попередньої обробки та редукції розмірності відгуків багатосенсорного масиву. Запропоновано критерій оцінки лінійної роздільності класів аналітів, який дозволяє вирішувати задачу вибору оптимального з точки зору класифікації методу обробки відгуків не вдаючись до перехресної перевірки з багатократним тренуванням класифікатора.

Ключові слова: багатоелементний масив сенсорів, кварцові мікроваги, лінійний класифікатор, редукція простору ознак, критерій роздільності класів.

Аннотация

ПРИМЕНЕНИЕ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ ДЛЯ АНАЛИЗА ИНФОРМАТИВНОСТИ СЕНСОРНЫХ МАССИВОВ ПРИ РЕГИСТРАЦИИ МОЛЕКУЛ ЛЕГКОЛЕТУЧИХ ВЕЩЕСТВ

Павлюченко А. С., Казанцева З. И., Кошец И. А., Ширшов Ю. М.

В работе рассмотрены некоторые подходы к улучшению качества классификации при распознавании химических образов за счет предварительной обработки и редукции размерности откликов многосенсорного массива. Предложен критерий оценки линейной разделимости классов аналитов, позволяющий решать задачу выбора оптимального с точки зрения классификации метода обработки откликов не прибегая к перекрестной проверке с многократным обучением классификатора.

Ключевые слова: многоэлементный массив сенсоров, кварцевые микровесы, линейный классификатор, редукция пространства признаков, критерий разделимости классов.

1. Introduction

Multielement chemical sensor arrays as an alternative to selective single sensor model in analytical equipment is a subject of intensive research during the last ten years. Main advantage of these devices is a possibility to perform qualitative analysis of the multicomponent mixtures without necessity to divide analytes into relevant and interfering.

As for today there exist several commercial production samples of multisensor systems intended mainly for qualitative or quantitative (with limited precision) express-analysis in the fields of food industry, medicine and manufacturing (see e. g. [1, 2]). However, despite the certain success of such instruments, problem of creation of the chemical sensors array based analytical system which would be effective, relatively cheap and easy to handle is very far from complete solution. In the process of development of actual instruments based on multielement sensor arrays a series of technical problems arises. These problems, while not being unsolvable in principle, may nullify the advantages of the approach if solved in nonoptimal way. Most significant of the problems are:

1) choice of the type and quantity of sensitive elements optimal for particular task or field of application;

2) ensuring stability of the sensitive elements responses (providing constant environmental conditions during the measurement or compensation of their drift);

3) selection and optimization of parameters for the measurement results processing algorithm.

The designer's task is to solve these problems in such way that it would allow to minimize the operating and maintenance cost and provide ease of handling while at the same time preserve simplicity of construction, relatively low prime cost and acceptable metrological performance.

Existing production samples are hard to qualify from this point of view since the detailed description of the employed engineering solutions, being a "know-how" of developer, is not a subject to publication. However even judging from a brief technical descriptions published by manufacturers it may be stated that these devices can be improved in many aspects.

The choice of the sensors type nowadays is usually performed empirically. Most frequently semiconductor sensors from metal-oxide or polymeric materials and composites changing their resistance during the contact with an analyzed gas are used [3]. Another frequently used class of sensors are the adsorptive sensors with various types of transducers (piezoelectric, optical etc.) utilizing thin layers of different organic molecules or molecular sieves as sorbents [4].

Various approaches to the solution of the above stated problems are possible. Problems of the sensors type and quantity selection along with the optimization of the measurement results processing algorithm parameters may be solved in two ways: either based on physical-chemical model of the processes taking place during the interaction of analyte with the sensor sensitive coating, or purely mathematically as problems of multi-objective optimization and multivariate statistical analysis.

Stability of the sensitive elements responses may be ensured either physically (through the specific device design) or by compensation of the influencing factors with the additional mathematical processing of the measurement results. In each case both meth-

ods derived from the physical-chemical model and not depending on such model may appear productive.

To this day, several partial models of interaction of the analyte vapor molecules with organic material were proposed. Among them are models depending on distribution of the analyte and sensitive layer molecules in soluted and gaseous phases [5], peculiarities of the sensitive layer molecular structure [6], two-stage adsorption onto sensor surface [7], relation of the interaction energies of central metal atom and peripheral substitutients in macrocyclic compounds [8]. Nevertheless, there is no adequate general physical-chemical model that would allow to completely describe interaction of the sensor sensitive element with the analyte.

Assumptions on which most of the existing models are based (binding only at the sensitive layer surface, monovalence and constant number of adsorption centers) in real systems are often not true. Besides that, it is usually necessary to take into account temporal drift of the sensitive layer parameters due to multiple influencing factors of presumably unknown nature. All this inevitably leads either to modification of the model with various heuristic constructs or to giving up the model at all.

In the further text we will consider some approaches to preprocessing of quartz crystal microbalance (QCM) responses in order to improve the discriminative ability of the sensor array. Response of a QCM sensor to exposition in analyzed atmosphere may be represented by continuous monotonic or nonmonotonic (depending on the sample preparation technique) function of time. In the case of monotonic function asymptotically approaching the steady state different characteristic parameters of the curve may be taken as single sensor response measure, e. g. maximum frequency shift [9], maximum value and the slopes of ascending and descending parts of the curve [10], integral value over the specifically chosen interval (so called *iV*-parameter) [11].

In this work we do not aim for definition of the general principles for optimization of the multielement sensor array; we will only study some practical methods of the measurement results processing as applied to the particular task. The discussed methods are based on statistical analysis of the sensors responses set and do not depend on any assumptions about the physical nature of these responses. Thus the obtained results may be also used (at least partially) when analyzing responses of the different type of sensors.

2. The concepts of chemical image and chemical images classification

The notion of "chemical image" is often encountered in various sources but it is rarely equipped by a rigorous definition. Appearance of the "chemical image" concept as an alternative to "chemical composition" is related to the attempts to simulate olfactory perception of the living organisms. It is obvious that mere enumeration of concentrations of the chemical mixture components is not enough to adequately represent the sense of odor. The description of complex physical-chemical structure of the analyzed mixture is necessary, reflecting the chemical composition, spatial disposition of the mixture particles and possibly other properties of the sample as well.

Multielement sensor array interacts with the analyte vapors and produces electrical signal at the output. Parameters of the signal may be measured and recorded with conventional measuring instruments. Measured values of these parameters may be represented by a multidimensional vector whose dimension is equal to the number of independent features of the signal, that is at the minimum is equal to the number of sensors in the array and may exceed it if we suppose that the output signal of each sensor is characterized by more than one parameter. It is assumed that obtained during the measurement multidimensional vector reflects specific properties of the analyte, that is at least a statistical relation exists between the "chemical image" of the analyte and the output signal of the sensor array. Existence of this relation is conditioned by the sensors cross-selectivity and its form is determined by peculiarities of the process of interaction between the analyte and the sensitive layer.

Unfortunately, lack of the model for interaction of analyte with the sensitive element does not allow to directly solve the inverse problem of sensor array synthesis depending on given specification of the "chemical image" mapping to output signal. In practice this problem is being solved by selection based on a posteriori analysis of information obtained from the test sensors set. Obviously, this set has to be superfluous for the sensors set found as a result of analysis to be close to optimal. The optimum criterion is determined by the field of application of the particular analytical instrument built with the multielement sensor array.

In actual applications set of possible "chemical images" of the samples subjected to analysis is usu-

ally known and limited. Thus the task of the sensor array output signal processing is reduced to matching of a certain label (class) to a multidimensional real vector, numerically describing this signal, that is, to construction of optimal in a certain sense classifier.

We will define *classifier* as a mapping of the following form:

$$\Phi(\mathbf{x}, X): R^n \to C \qquad (1)$$

where $\mathbf{x}$ is a n-dimensional real random vector with unknown in general case distribution; $X \subset R^n$ is a training set (countable and finite subset of R^n for which the mapping $X \to C$ is defined explicitly); C is countable (and, as a rule, finite) set of classes. Construction of the mapping Φ consists in certain estimation of the form and parameters of the $\mathbf{x}$ distribution within each class based on the training set X and construction of the classification rule depending on the found estimate. Since (1) maps noncountable and, generally speking, infinite set of real vectors to the countable and finite set of classes then each class i from the set C corresponds to an area $X_i \subset R^n$. In practice $\mathbf{x}$ may take values only from some subset of R^n and the probability density functions of $\mathbf{x}$ within each class are unimodal; thus as a rule X_i are bounded. Obviously, a "good" classifier must provide disjointness of the areas X_i and each area X_i must contain maximum of the $\mathbf{x}$ distribution density within the class i.

Actually, not all of the $\mathbf{x}$ vector elements are equally informative, both due to the sensor array superfluity and due to the fact that certain functions of several elements of the multidimensional random vector may be more informative than each of this elements taken separately [13]. Thus preprocessing of the sensor array output signal that leads to dimensionality reduction may improve classifier performance.

If the mapping (1) can be represented in form

$$\Phi(f(\mathbf{x}), X): R^n \to C \qquad (2)$$

where $f(\cdot): R^n \to R^m, m < n$, then we will call such classifier a classifier with the feature space reduction.

Feature space reduction is useful also in the sense that the decrease of processed data amount allows to simplify processing algorithms and reduce processing time as well as requirements to the system memory amount, which may be significant when developing stand-alone analytical systems with built-in microprocessor units.

Choice of the form of function $f(\cdot)$, however, possesses the same difficulties that the process of sensor array optimization in general: there is no uniform formalized approach to the solution of this problem, so the selection based on a posteriori classification performance estimates has to be employed.

In practice situation gets even more complicated by the fact that the $\mathbf{x}$ distribution parameters are not constant because of change of environmental conditions during the measurement and the sensors contamination and ageing. This leads to necessity of periodical renewal of the X set and reconstruction of the Φ mapping according to the new training set. This approach has obvious disadvantages and in many applications may be unacceptable. The simplest method of compensation of the parameters drift is normalization of the classifier input vectors (as well as vectors in the training set), assuming that the parameters of $\mathbf{x}$ distribution linearly depend on influencing factors. This, however, is not always true; besides, normalization, while not being a reduction procedure is the sense of (2), nevertheless leads to reduction of the $\mathbf{x}$ entropy and thus in some cases may lead to the loss of relevant information, so it has to be used carefully.

The methods of features drift compensation based on the systems identification theory are developed [14, 15]. They consist in construction of a dynamic model of the sensor array and adaptive correction of the model parameters according to the new data incoming from the actual sensors. Classification of the samples is performed by means of comparison of the computed model responses with the measured response of the real sensor array. Such approach, however, results in a large amount of computations that need to be performed during the model construction and afterwards during the each parameters correction procedure, which has to take place often enough (otherwise the sensors parameters may significantly change as compared to their values during the previous correction which in its turn may lead to the fault of the optimization algorithm). In view of the aforesaid, these methods seem to be too complicated for the wide application (at least in the portable analytical systems with the built-in data processing facilities).

Classifier performance, naturally, depends on the classification rule itself as well. Various types of classifiers are mentioned in the literature as applied to the analysis of multielement chemical sensor arrays responses [16]. Most popular of them are Prin-

cipal Component Analysis (PCA) [17, 18] and multilayer Artificial Neural Networks (ANN) trained by back propagation of error [19]. However, despite their wide spread occurrence, these methods do not provide complete solution of the classification problem. Indeed, PCA itself, strictly speaking, is not a classification rule, that is it does not allow to automatically classify unknown input vectors, though it may prove useful as a feature space reduction or preliminary data exploration method.

As for the multilayer ANN, their main advantage is an ability to form a classification rule separating the input vectors space into areas bounded by nonlinear (and even not convex) surfaces. However this may be significant only in the case when we have a reason to suppose that the $\mathbf{x}$ distribution within one or several classes significantly differs from gaussian and its probability density function has several extrema. Obviously, the vectors with such distribution are unlikely to appear among the sensor array responses (since we suppose that responses of a single sensor towards the same analyte can not significantly change from measurement to measurement, and slight changes are conditioned by fluctuations of the environment parameters — temperature, humidity, analyte concentration etc.) and even if they do, such phenomenon should be considered an anomaly and evidence of defect in the system design rather than actual reflection of peculiarities in the analyzed data structure.

When dealing with the training sets of small size (which is usually the case in the studied field because of impossibility to perform multiple measurements in a short time) nonlinearity of classifier may easily lead to overtraining. From the computational point of view both PCA and ANN are implemented in the form of rather complex iterative algorithms and possessing potential numeric instability. On the other hand, known results of performance evaluation for ANN as applied to classification of chemical images [20, 21] do not show their explicit superiority in comparison with simpler classifiers, including linear ones. Although in some cases ANN-based classifiers may have somewhat better performance than the more simple ones, their utilization for estimation of classes separability during the optimization of chemical sensors array, however, results in significant increase of machine time consumption. Besides, optimization of the classifier architecture itself may become a non-trivial task in this case.

A *linear classifier* is a mapping (1), for which the following rule holds:

$$\Phi(\mathbf{x}, X) = i : \mathbf{n}_{ij}(X) \cdot \mathbf{x} < \beta_{ij}(X) \quad \forall j \in C, j \neq i \quad (3)$$

where $\mathbf{n}_{ij}$ is a normal vector of a hyperplane separating areas of classes i and j; β_{ij} is a threshold determining position of the separating hyperplane relative to the coordinates origin. $\mathbf{n}$ and β determine the position of the plane separating two classes and both are functions of a training set. The actual form of these functions is determined by a classifier type.

In practical tasks of chemical images recognition, as was already said above, we usually have to deal with the small-size training sets, when the number of samples in each class is less or just slightly greater than the vector dimensionality. In this case it is difficult or impossible to soundly estimate parameters of vector distribution in each class, let alone the form of the distribution itself. Thus we have to utilize classifiers with a classification rule depending on minimum statistical information. However, even such simple means allow to achieve acceptable results. Optimization of the classification process in this case is reduced to selection of the best in a certain sense procedure of preprocessing the $\mathbf{x}$ vector (including, possibly, a feature space reduction procedure as well), since the classifier itself due to its simplicity does not have enough parameters which could be varied for optimization purpose.

In this work we will use a "nearest mean vector" classifier with euclidean metric. Let us define the formal classification rule. To do this we substitute $\mathbf{n}$ and β in (3) with appropriate expressions: $\mathbf{n}_{ij} = \mathbf{m}_j - \mathbf{m}_i$; $\beta_{ij} = (\mathbf{m}_j - \mathbf{m}_i) \cdot (\mathbf{m}_j + \mathbf{m}_i)/2$ (the normalizing factor is omitted). Hence

$$\Phi(\mathbf{x}, X) = i : (\mathbf{x} - (\mathbf{m}_j + \mathbf{m}_i)/2) \cdot (\mathbf{m}_j - \mathbf{m}_i) < 0 \quad \forall j \in C, j \neq i \quad (4)$$

where $\mathbf{m}_j$, $\mathbf{m}_i$ are the expectation values of vectors within j-th and i-th classes respectively. In practice expectations are replaced by mean values calculated over the training set.

It is easy to see that the rule (4) is nothing else than the maximum likelihood criterion for the Fischer's model with a diagonal covariance matrix with equal diagonal elements [see, e. g., 22, p. 50]. This criterion may be considered natural in assumption of gaussian distribution of $\mathbf{x}$ within the each class and taking into account impossibility to estimate the actual covariance matrix.

Performance of the classifier (4) depends only on the training set, that is, assuming the correct meas-

urement technique, only on measurement results preprocessing method. To construct an optimal classifier, let us introduce the criterion of separability for classes *j* and *i*:

$$s_{ij} = \frac{|\mathbf{m}_j - \mathbf{m}_i|}{|\mathbf{m}_j - \mathbf{m}_i| + \varsigma} \quad (5)$$

where $\mathbf{m}_j$ and $\mathbf{m}_i$ have the same meaning that in (4), $|\cdot|$ is the euclidean norm and ς is a sum of absolute values of projections of the standard deviations within classes *j* and *i* to the vector $\mathbf{m}_j - \mathbf{m}_i$. In practice standard deviations are replaced by their estimates computed over the training set.

Criterion (5) equals 1 if the distance between mean vectors is non-zero and standard deviations within classes are zero, and it decreases as the variance within at least one of the classes increase. It is obvious that the classes may be considered well-separable if $s_{ij} > 0.5$ (if we suppose that the absolute values of projection of standard deviations within classes *i* and *j* to the vector $\mathbf{m}_j - \mathbf{m}_i$ are the "effective radii" of the classes then the case of $s_{ij} = 0.5$ corresponds to the intersection of hyperspheres of the corresponding radii in a single point; when $s_{ij} > 0.5$ hyperspheres do not intersect).

Criterion (5) may be computed for each pair of classes from the whole ensemble of classes defined for particular task. To evaluate the classification performance in general, however, we need to introduce some integral characteristic, based upon the values of s_{ij}. We will use the geometric mean for this purpose:

$$S_I = \sqrt[k]{\prod_C s_{ij}} \quad (6)$$

where $k = \mathrm{card}(C)$ is a number of classes in C.

Geometric mean appears to be a suitable integral criterion since it is known that for asymmetric distributions it is a better estimate of the central tendency than the arithmetic mean. In other words, presence of the small values among the s_{ij} leads to a stronger "penalty" for geometric mean used as an integral criterion. It is possible however that better forms of the integral separability estimates based on criterion (5) exist; this question requires further investigation.

In the following text we will study the influence of various methods of sensor array responses preprocessing and sensors set selection on separability of the obtained set of chemical images classes utilizing criterion (6) as well as correlation of criterion (6) with the estimate of classification performance evaluated by the leave-one-out cross-validation technique (that is, excluding one of the vectors from the training set, construction of classifier using the remaining vectors and testing it with the excluded vector taken as input; the procedure is repeated for all vectors in the training set).

3. Experimental

The experimental database was obtained on an injection-type gas chamber made from Teflon and containing eight quartz resonators with the sensitive layers formed by thin (about 200 nm) films of thermally evaporated substituted calixarenes. The gas flow system was alternately connected either with the injector syringe (in measurement mode) or with the membrane pump (in cleaning mode) by means of two-position valve. Cleaning of sensors after the each measurement was performed by blowing with the desiccated air until restoration of the base frequency. More detailed description of the experimental installation may be found in [23].

Coatings of derived calixarenes with 4, 6 and 8 phenol rings and different functional groups synthesized in the Institute of Organic Chemistry (Kiev, Ukraine) were used as sensitive layers. It is known [23] that spatial structure of calix[n]arenes and calix[4]resorcinarenes molecules makes up cavities which due to their geometry become selective receptors for the neutral organic molecules, especially for benzene derivatives. 14 different types of sensitive coatings were synthesized. Molecular structure of the similar materials is described in [23, 24].

Seven volatile organic compounds vapors were used as analytes (carbon tetrachloride, benzene, chloroform, methylene chloride, dichlorethane, toluene and xylene) at concentration of approximately 1000 ppm.

Temporal change of the crystal resonance frequency relative to the base one was considered a sensor response. From 4 trough 7 responses towards each of the analytes were obtained. Eight calixarene coatings was selected for experiments — tetrapropylcalix [4]resorcinarene (CA2), tio *tert*-butylcalix [4]arene (CA6), tetraformyltetrapropoxycalix [4]arene (CA7), propoxycalix[6]arene (CA10), *tert*-butylcalix[6]arene (CA11), *tert*–butylcalix[8]arene (CA12), octakis-diethoxyphosphoryloxycalix [8]arene (CA13) and diethoxyphosphoryloxycalix [4]arene (CA14). Typical sensor array response is shown in fig. 1.

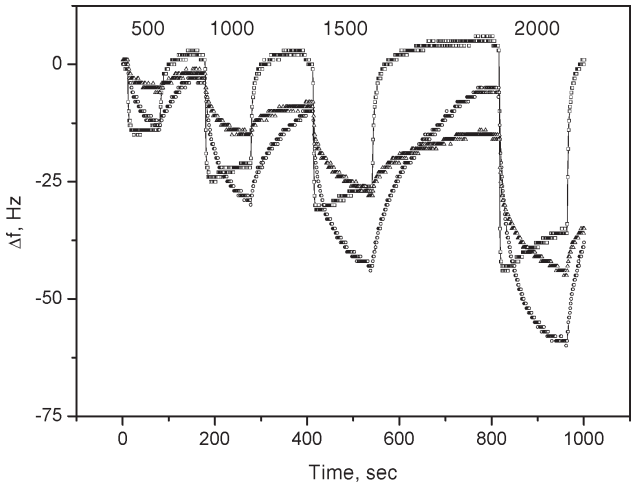

Fig. 1 Typical sensor array response to injection of the chloroform vapor (responses of only three sensors are shown). Numbers at the top of the graph denote concentration of the injected analyte.

4. Results and processing

For recognition tests and further evaluation of the studied analytes classification performance a database was formed, where every response curve was represented by 59 sample points (according to the time length of the shortest response) taken with a period of 1 second starting from the analyte injection moment. All curves were corrected by subtraction of averaged measured value of the sensor crystal base frequency and smoothed by a third order moving average filter. Responses of some sensors towards particular analytes showed increase of the resonance frequency instead of usual decrease, probably related to desaturation of the water molecules adsorbed on the sensitive layer surface. Such samples were set to zero.

Examples of the response curves are shown in fig. 2. It can be seen that some response values are equal to zero for all sample points of the curve. However, these responses intentionally were not excluded from the further analysis.

In order to evaluate reproducibility of sensor responses the standard deviations for each sample of the response curve starting from tenth one (to avoid sample values close to zero) were computed within the each analyte class. Maximum sensor-wise values of standard deviations reduced to the corresponding mean values were: 0. 28, 0. 43, 0. 26, 0. 17, 0. 54, 0. 31, 0. 21, 0. 09 for responses to carbon tetrachloride; 0. 22, 0. 00, 0. 22, 0. 50, 0. 37, 0. 34, 0. 23, 0. 30 for responses to benzene (the second sensor always reacted to this analyte with resonance frequency increase); 0. 10, 0. 14, 0. 14, 0. 16, 0. 26, 0. 16, 0. 13, 0. 10 for responses to chloroform; 0. 37, 0. 30, 0. 14, 0. 20, 0. 05, 0. 06, 0. 09, 0. 08 for responses to methylene chloride; 0. 05, 0. 40, 0. 09, 0. 24, 0. 10, 0. 10, 0. 06, 0. 09 for responses to dichlorethane; 0. 42, 0. 56, 0. 70, 0. 25, 0. 81, 0. 64, 0. 27, 0. 17 for responses to toluene; 0. 39, 2. 00, 0. 42, 0. 29, 0. 42, 0. 40, 0. 33, 0. 41 for responses to xylene (the second sensor is not sensitive towards this analyte). The general tendency is the response deviation increase as the amplitude of response decreases, that is, better response stability is proper to the senors with higher sensitivity towards a given analyte.

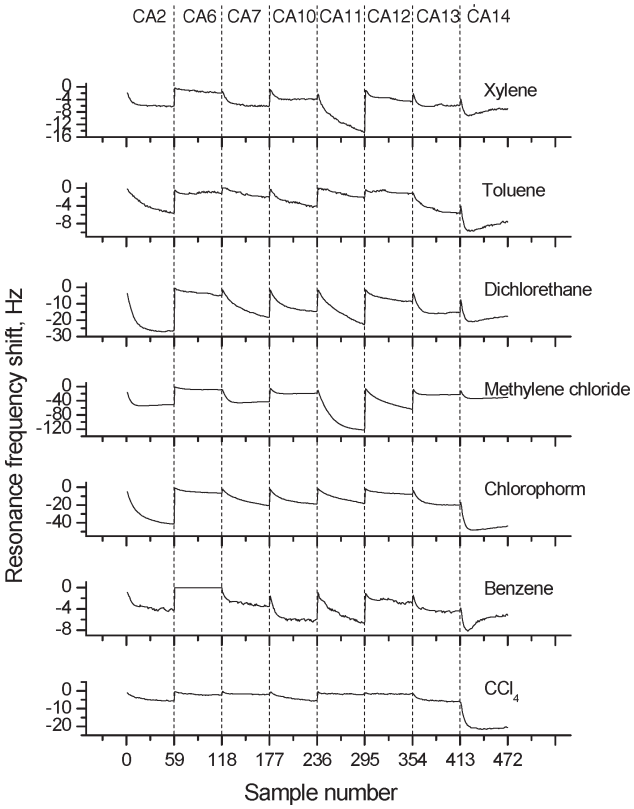

Fig. 2. Examples of sensor array response curves for the seven different analyte vapors (composite curves used for classification).

Obtained curves were treated with three different preprocessing methods with dimensionality reduction in order to choose the optimal one. Two of the methods (averaging and extremum search) were chosen as the most frequently encountered in practical implementations of the similar sensor systems. The third method, utilizing statistical properties of responses, was chosen to evaluate potential possibility to improve classification performance with a data processing which does not depend on the physical nature of responses (similar approach is used in [27]).

Thus from the obtained response curves four data sets were formed:

1) without any special preprocessing — each sensor response represented by 59 sample points of the frequency change curve, sensor array response is represented by 472 samples of the sequentially attached sensor responses.

2) Maximum shift of the crystal resonance frequency relative to the base one is used as a determinative feature of sensor response.

3) Average value of the resonance frequency shift taken over all of the response curve sample points is used as a determinative feature of sensor response.

4) Resonance frequency shift value in the fixed sample point of the response curve is used as a determinative feature of sensor response. The number *i* of the sample point was determined for each sensor according to the maximum of $\frac{\sigma_{\text{inter}}}{\sum_C \sigma_{\text{intra}}}$ criterion, where σ_{inter} is an estimate of the standard deviation of the mean response values of the each class, σ_{intra} is an estimate of the standard deviation of response within the each class, both computed for each sample point of the response curve. Positions of selected points on the averaged response curves are shown in fig. 3.

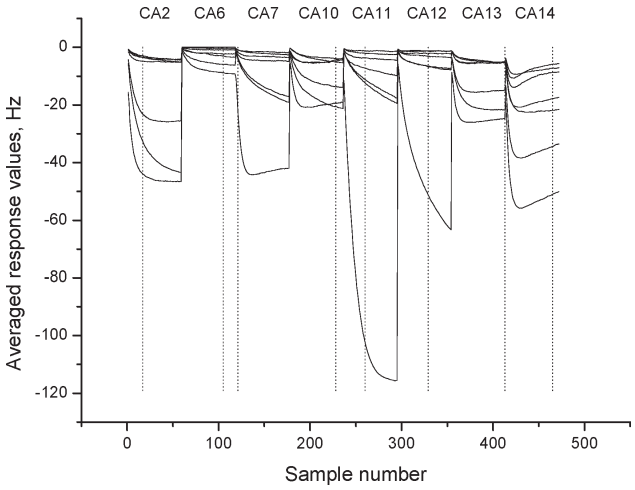

Fig. 3. Averaged response curves of eight sensors towards the vapors of seven different analytes. Dot lines show positions of the samples selected according to the standard deviations ratio criterion to form the fourth group of data sets.

From every of the four described data sets the subsets were obtained by sequential decrease of number of sensors and exhaustive enumeration of all possible combinations of responses for each number of sensors from 1 through 7. Totally 1020 different data sets were formed, consisting of the four groups (corresponding to data processing method) each containing 255 sets. In the first group (non-reduced responses) each response was represented by 59 sample values and dimensionality of the feature vector thus was $59 \cdot N$, where N is a number of sensors; in the three remaining groups each response was represented by a single feature obtained as a reduction result, and dimentionality of the feature vector was equal to the number of sensors. For every data set value of the criterion (6) was computed in order to determine an optimal sensors combination for the each response processing method. Moreover, each of the data sets was used as a training set for leave-one-out cross-validation of the classifier (4).

Fig. 4 shows dependence between the criterion (6) and classification performance determined by cross-validation for the four data groups. Figures 5 and 6 show dependence of criterion (6) and classification performance determined by cross-validation on the number of sensors in the set (for these graphs the best sensor sets were taken from the each group of sets with equal number of sensors). Table 1 shows the best (in the sense of criterion (6) and cross-validation result) sensor sets.

5. Discussion

As can be seen from fig. 4, criterions S_1 and k_+ / k (where k_+ is a number of correctly recognized during the leave-one-out cross-validation classes, k is the overall number of classes) are correlated, but the sensor sets optimal in the sense of first and second criterions do not match. This, however, does not necessarily mean inefficiency of the S_1 criterion. Both S_1 and k_+ / k are random variables, and it is known [25] that the k_+ / k estimate obtained with leave-one-out method has greater variance compared to the classification error estimates determined by other methods. Statistical properties of the criterion (6) of course demand for further study. Nevertheless, the fact that optimal in the sense of criterion (6) sensor set is practically identical for all considered processing methods allows to make a preliminary conclusion of possibility to utilize this criterion in solution of the problem of chemical images recognition process optimization.

It is worth to note, that sensor 2 belongs to the optimal set despite its abnormal reaction towards the vapors of some analytes, as a result of which corresponding response curve samples were set to zero.

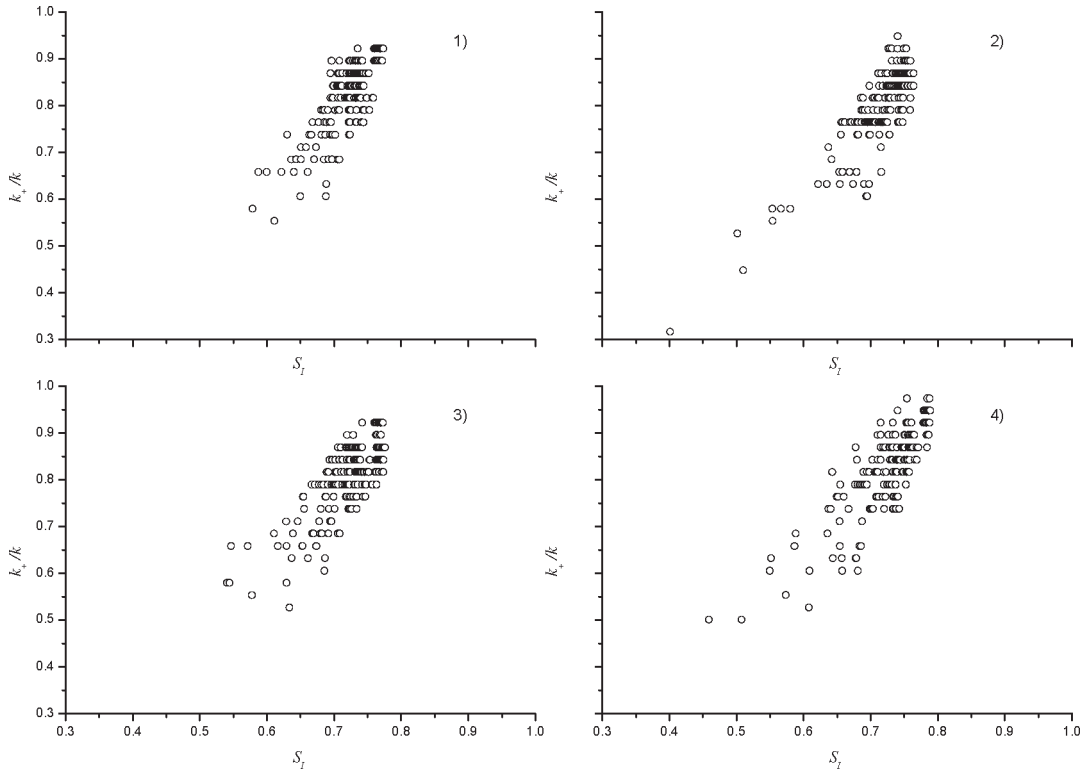

Fig. 4. Dependence between the values of integral separability criterion and classification performance determined by leave-one-out cross-validation for the four data sets: 1) all sample points of the response curve; 2) maximum resonance frequency shift; 3) average resonance frequency shift; 4) resonance frequency shift in the point defined by standard deviations ratio criterion.

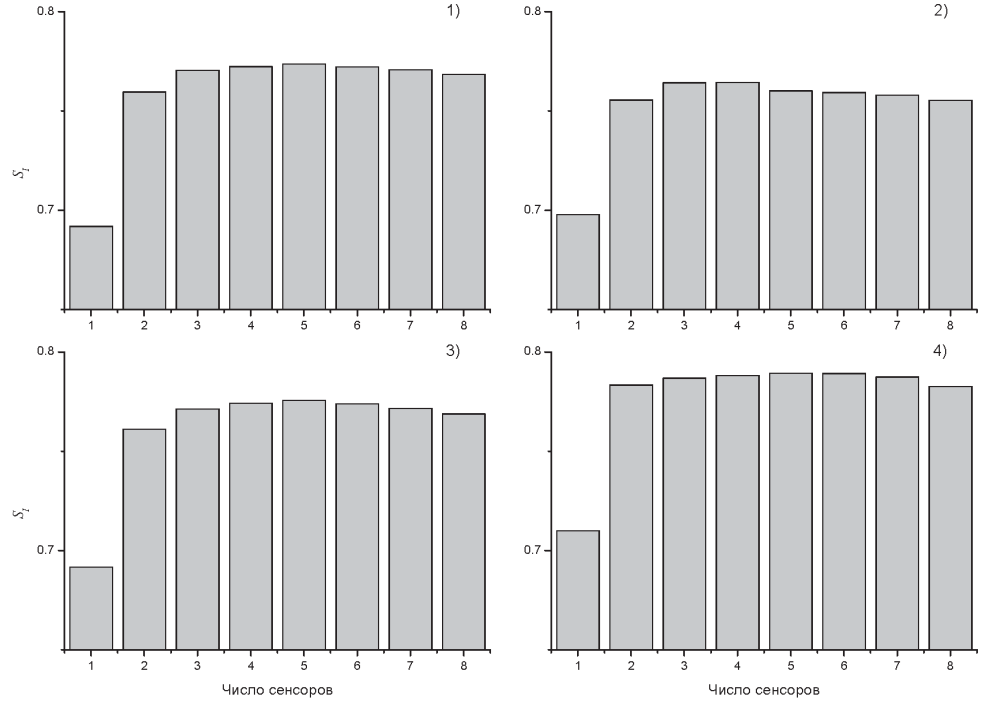

Fig. 5. Dependence of criterion S_1 on the number of sensors for the four data sets: 1) all sample points of the response curve; 2) maximum resonance frequency shift; 3) average resonance frequency shift; 4) resonance frequency shift in the point defined by standard deviations ratio criterion.

Table 1

Processing method	Best sensors set according to S_I criterion	Maximum value of the S_I criterion	Best sensors sets according to k_+/k criterion	Maximum value of the k_+/k criterion
1	1-2-5-6-8	0.773566	2-6 ... 1-2-3-4-5-6-7-8*	0.921
2	1-2-5-6	0.764288	1-2-3-4-6	0.947
3	1-2-5-6-8	0.775599	2-6 ... 1-2-3-4-5-6-7-8**	0.921
4	1-2-5-6-8	0.789171	2-6 2-5-8 2-5-6-7-8	0.974

* Totally 53 sets
** Totally 25 sets

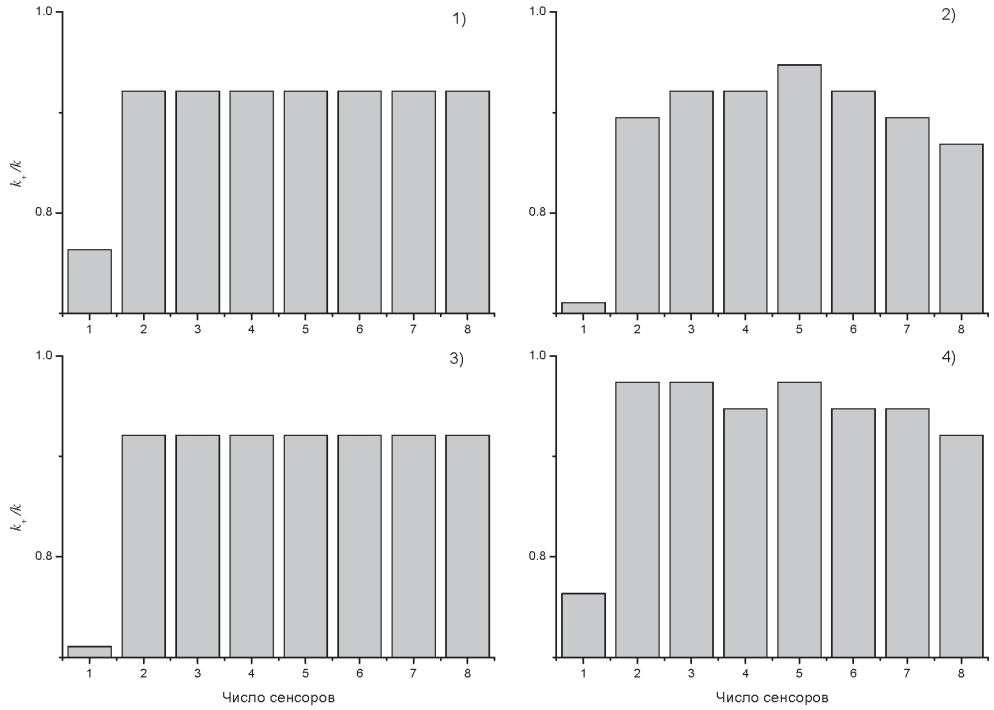

Fig. 6. Dependence of classification performance determined by leave-one-out cross-validation on the number of sensors for the four data sets: 1) all sample points of the response curve; 2) maximum resonance frequency shift; 3) average resonance frequency shift; 4) resonance frequency shift in the point defined by standard deviations ratio criterion.

Performed analysis also allows to state that the set of eight sensors is superfluous for the studied recognition task. As can be seen from figures 5 and 6 optimum number of sensors for recognition of seven analyte classes is either four or five. General form of the dependencies shown in figures 5, 6 coincides with the results obtained in [26] for the sensors of different type, though the particular optimal number of sensors and composition of the sensor set apparently depend on number of classes and analytes nature. For example, if we split the data set obtained with processing method 4) (providing the best values of classification performance criterions) to the two subsets, first containing the data for first three analyte classes (carbon tetrachloride, benzene, chloroform) and second containing the data for the remaining classes (methylene chloride, dichlorethane, toluene, xylene) then the optimal in the sense of criterion (6) sensor sets will be 1-2 and 4-7 respectively and the value of S_I criterion will amount to 0. 882543 for the first set and 0. 523539 for the second set.

As for the methods of characteristic features ex-

traction from the sensor array response, methods 2) and 3) though often utilized in practice do not provide the best possible result, moreover, it may be even worse than the absence of any extraction at all. Method 4), while not utilizing any information of actual nature of responses and being based only on their statistical features estimate, nevertheless allows to improve classification performance.

6. Conclusion

Performed analysis of responses of QCM sensors array towards the vapors of volatile compounds allows to state that even initially obtained with the beforehand chosen response processing method high classification performance does not necessarily imply optimality of either the sensors set or the method itself. Moreover, classification performance may significantly vary for the different analyte classes sets even with the same set of sensors. Hence the choice of an optimal sensors set and an optimal method of processing of the sensor array output signals have to be performed individually for each set of the analyte classes, and the optimality criterion must reflect characteristic peculiarities of the task being solved.

Proposed in this work integral criterion for the linear separability of analyte classes allows to evaluate, at least preliminary, the effectiveness of the chosen responses processing as applied to the linear classification performance improvement without the need to resort to time-consuming cross-validation techniques requiring multiple retraining of classifier. It is shown that with the help of relatively simple in computational sense separability evaluation procedure it is possible to improve the chemical images classification performance with simultaneous decrease of number of sensors in the array and minimization of number of independent parameters for the each sensor response, which is especially important for development of compact stand-alone analytical systems for qualitative express-analysis of the gaseous medium samples.

References

1. http://sensors-transducers. globalspec. com/SpecSearch/AllSuppliersByProductArea?LSArea=2003
2. http://www. technobiochip. com
3. 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 //Chemical Review 100 (2000), pp. 2595–2626
4. C. Di Natale, R. Paolesse, A. Macagnano, A. Mantini, C. Goletti, A. Amico. Characterisation and design of porphyrin based broad selectivity chemical sensors for electronic nose applications // Sensors and Actuators B 52 (1998), pp. 162–168
5. J. V. Grate, M. H. Abraham. Solubility interactions and the design of chemically selective sorbent coatings for sensors and sensor arrays //Sensors and Actuators B 2 (1991), pp. 85–111
6. J. Hartmann, P. Hauptmann, S. Levi, E. Dalcanale. Chemical sensing with cavitands: influence of cavity shape and dimensions on the detection of solvent vapors //Sensors and Actuators B 35-36 (1996), pp. 154–157
7. M. Horn. A new theory of adsorption for the quantitative description of gas sensors //Sensors and Actuators B 26-27 (1995), pp 217–219
8. C. Di Natale, R. Paolesse, A. Macagnano, A. Mantini and A. D'Amico. Selectivity tailoring in porphirins-based QMB sensors for volatile compounds //Proc. Of Eurosensors XIII, Hague 1999, pp. 219220.
9. J. W. Gardner, T. C. Pearce, Friel et al. A multisensor system for beer flavour monitoring using an array of conducting polymers and predictive classifiers. //Sensors and Actuators B 18-19 (1994), pp. 240–243
10. T. D. Gibson, O. C. Prosser, J. Peace, J. N. Hulbert. From laboratory to factory: the use of Electron Nose technology in on0line monitoring // Proc. of Eurosensors XII, Southhampton 1998, pp. 1131–1134.
11. B. A. Snopok, I. V. Kruglenko. Multisensor system for chemical analysis: state-of-the-art in Electronic Nose technology and new trends in mashine olfaction //Thin Solid Films 418 (2002), pp. 21–41.
12. Wolfgang Gцpel. Chemical imaging: I. Concepts and visions for electronic and bioelectronic noses // Sensors and Actuators B 52 (1998), pp. 125–142
13. Isabelle Guyon, Andrй Elisseeff. An Introduction to Variable and Feature Selection //Journal of Machine Learning Research 3 (2003), pp. 1157-1182
14. Martin Holmberg, Ingemar Lundstrцm, Fredrik Winquist, Fabrizio A. M. Davide, Corrado Di Natale, Arnaldo D'Amico. Drift counteraction for an electronic nose //Sensors and Actuators B 35–36 (1996), pp. 528–535
15. Martin Holmberg, Fabrizio A. M. Davide, Corrado Di Natale, Arnaldo D'Amico, Fredrik Winquist, Ingemar Lundstrцm. Drift counteraction in odour recognition applications: lifelong calibration method //Sensors and Actuators B 42 (1997), pp. 185–194
16. P. C. Jurs, G. A. Bakken, H. E. McClelland.

Computational Methods for the Analysis of Chemical Sensor Array Data from Volatile Analytes // Chemical Review 100 (2000), pp. 2649–2678
17. K. Nakamura, T. Nakamoto, T. Moriizumi. Classification and evaluation of sensing films for QCM odor sensors by steady-state sensor response measurement //Sensors and Actuators B 69 (2000), pp. 295–301
18. Corrado Di Natale, Antonella Macagnano, Sara Nardis, Roberto Paolesse, Christian Falconi, Emanuella Proietti, Pietro Siciliano, Roberto Rella, Antonella Taurino, Arnaldo D'Amico. Comparison and integration of arrays of quartz resonators and metal-oxide semiconductor chemoresistors in the quality evaluation of olive oils //Sensors and Actuators B 69 (2000), pp. 303–309
19. H. Nanto, K. Kondo, M. Habara, Y. Douguchi, R. I. Waite, H. Nakazumi. Idetintification of aromas from alcohols using a Japanese-lacquer-film-coated quartz resonator gas sensor in conjunction with pattern recognition analysis //Sensors and Actuators B 35–36 (1996), pp. 183–186
20. Snopok B. A., Kruglenko I. V., Shirshov Yu. M, Reznik A. M., Nowicki D. W., Dekhtyarenko A. K. Computational Selectivity of Chemical Arrays: Associative Memories Algorithms as Effective Classifier for Electronic Nose Applications //Sensors for Environmental Control (Ed. P. Siciliano), World Scientific Publishing Co. Pte. Ltd., 2003, pp. 239–243
21. Manuele Bisego, Gino Tessari, Giampietro Tecchiolli, Marco Bettinelli. A comparative analysis of basic pattern recognition techniques for the development of small size electronic nose //Sensors and Actuators B 85 (2002), pp. 137–144
22. С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989
23. V. I. Kalchenko, I. A. Koshets, E. P. Matsas, O. N. Kopylov, A. Solovyov, Z. I. Kazantseva, Yu. M. Shirshov. Calixarene-based QCM Sensors Array and Its Response to Volatile Organic Vapours //Materials Science Vol. 20, No. 3 2002
24. И. А. Кошец, З. И. Казанцева, Ю. М. Ширшов, С. А. Черенок, В. И. Кальченко. Каликсарены как чувствительные слои для газовых сенсоров на основе кварцевого микробаланса // Оптоэлектроника и полупроводниковая техника, 2003, вып. 38 (в печати)
25. Anil K. Jain, Jianchang Mao. Statistical Pattern Recognition: A Review //IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, No. 1, January 2000, pp. 4–35
26. Michael C. Burl, Brian C. Sisk, Thomas P. Vaid, Nathan S. Lewis. Classification performance of carbon black-polymer composite vapor detector arrays as a function of array size and detector composition //Sensors and Actuators B 87 (2002), pp. 130–149
27. I. V. Kruglenko, B. A. Snopok, Yu. M. Shirshov, F. J. Rowell. Multisensor systems for gas analysis: optimization of the array for the classification of the pharmaceutical products //Semiconductor Physics, Quantum Electronics & Optoelectronics 2004 (in press)