www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

MINING DATA OF NOISY SIGNAL PATTERNS IN RECOGNITION OF GASOLINE BIO-BASED ADDITIVES USING ELECTRONIC NOSE

Stanisław Osowski[1,2], Krzysztof Siwek[1]

1) Warsaw University of Technology, Faculty of Electrical Engineering, Pl. Politechniki 1, 00-661 Warsaw, Poland (✉ sto@iem.pw.edu.pl, +48 22 234 7235, ksiwek@iem.pw.edu.pl)

2) Military University of Technology, Faculty of Electronic Engineering, Gen. S. Kaliskiego 1, 00-908 Warsaw, Poland

Abstract

The paper analyses the distorted data of an electronic nose in recognizing the gasoline bio-based additives. Different tools of data mining, such as the methods of data clustering, principal component analysis, wavelet transformation, support vector machine and random forest of decision trees are applied. A special stress is put on the robustness of signal processing systems to the noise distorting the registered sensor signals. A special de-noising procedure based on application of discrete wavelet transformation has been proposed. This procedure enables to reduce the error rate of recognition in a significant way. The numerical results of experiments devoted to the recognition of different blends of gasoline have shown the superiority of support vector machine in a noisy environment of measurement.

Keywords: data mining, electronic nose, gasoline blends, random forest, support vector machine, wavelet de-noising.

1. Introduction

The subject of the paper is studying the properties of sensor signals distorted by the random noise in an electronic nose. The noise present in the nose measurement may be a result of thermal fluctuations, power drift, sensor instability or chemical interfering agents, *etc*. The additive Gaussian model of noise affecting the semiconductor sensors is usually assumed in considerations and this model has been examined in the paper. The noise makes measurements not-repeatable, so analyzing the effect of this distortion is significant from a practical point of view.

In the paper this distortion is statistically analyzed on the examples of gasoline blends built on the basis of different bio-products, such as ethanol, *methyl tertiary butyl ether* (MTBE*), ethyl tertiary butyl ether* (ETBE) and benzene. These supplements are used as fuel oxygenates to increase the octane index (to replace the banned tetraethyl lead) or to raise the oxygen content in gasoline. The addition of the bio-products changes the odor of a blend and thus enables to recognize its type [1, 2]. An electronic nose, being able to recognize different types of blends, may find practical application in building a low-cost, very fast and accurate measuring instrument for recognizing the gasoline blends. This is an important problem in checking the quality and parameters of gasoline blends sold in petrol stations.

Different types of sensors are used in an electronic nose. A good review of them is presented in the papers [3] and [4]. In our application we have used an array of semiconductor sensors, very popular due to their wide availability, ease of use and relatively low cost. They form the heart of an electronic nose and respond with a signal pattern according to the odor of each gasoline blend. Analysis of these signals can answer questions regarding recognition of the supplemented bio-products directly on the basis of the blend odor [2, 3, 5]. One of the most

Article history: received on Aug. 11, 2016; accepted on Jan. 14, 2017; available online on Mar. 10, 2017; DOI: 10.1515/mms-2017-0015.

important problems in this task is assuring the robustness of a nose to the disturbed measurement signals in connection with the number of applied sensors. The distortion of sensor signals is recognized as small and non-deterministic temporal variations. It is the result of changing the working conditions of the semiconductor sensors following the changes of the environmental parameters (temperature, humidity, pressure), as well as sensor circuit instability, which occur in the measurement process [6, 7].

There are some papers studying the effect of noise and drift in the nose measurement process and methods of its reduction [6–10]. They are based on application of compression techniques and apply the methods of principal component analysis, wavelet transformation or singular value decomposition. Most of the papers concerning the pattern recognition in an electronic nose deal only with the actually measured signals, not analyzing the effect of their additional distortion by noise [1–3]. The most often used techniques apply the neural networks, the support vector machine, the principal component analysis, the linear discriminant analysis, *etc.*

The aim of the paper is to study the statistical effect of noise disturbing the semiconductor sensor signals in an electronic nose, used in the process of recognition of the gasoline blends.

The problem of a long time drift is not considered. Different levels of white Gaussian noise are applied and a few aspects of the problem are considered:

- The unsupervised analysis of data distribution at different levels of distortion; it is based on clustering the measured samples and analyzing their changes caused by noise.
- The supervised analysis directed to pattern recognition and classification, in order to determine the robustness of system to noise.
- Improvement of the pattern recognition accuracy by applying a de-noising procedure to the noisy sensor signals.

The paper compares the performance of two most efficient classification systems: the random forest of decision trees [11] and the support vector machine [12]. The random forest proposed by Breiman in 2001 is based on a learning strategy called "ensemble learning" with generating many classifiers and aggregating their results according to the majority voting rule. The random forest can be directly applied to solve the recognition and classification tasks. It also provides a measure of significance of a particular sensor in the measurement process. In this respect it is a very useful tool, more universal than most of the known solutions of signal processing in an electronic nose, such as the linear discriminant analysis, *principal component analysis* (PCA), neuro-fuzzy systems or neural networks [13–17]. On the other hand, the support vector machine was created as a classification tool significantly resistant to noise distorting the input data [12, 17, 18]. Both methods are compared for the data distorted by the random noise with a normal distribution, zero mean and different variance values.

An important aspect considered in the paper is reduction of the noise contaminating the sensor signals. The discrete wavelet transform, decomposing sensor signal patterns into wavelet components of different resolution, is studied. The noise influence is reduced by cutting the detailed signals of their lowest values. Thanks to this the final classification accuracy of patterns is achieved. The results of numerical experiments have shown a high efficiency of this strategy in improving the recognition of patterns formed by the sensor signals.

2. The electronic nose measurements

Recognition of the gasoline blends on the basis of their odor exploits the fact that the blends are associated with different odors resulting from their chemical composition. Analysis of the odor is a complex issue because of a heterogeneous nature of gasoline. Application of an electronic nose and artificial intelligence methods seems to be an efficient way of analyzing the odor [2–4]. The patterns of signals of the vapor-sensitive detectors are processed in this approach and associated with different types (classes) of gasoline blends.

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

Many different techniques of signal processing have been employed in an electronic nose. They include: the *principal component analysis* (PCA) and linear discriminant analysis [3], self-organizing maps [16], the *k*-nearest neighbor algorithm [15], neuro-fuzzy systems [14], different types of neural networks [1, 3], the support vector machine [2, 17, 18]. Recently, the random forest has been also tried in recognition of orange beverage and Chinese vinegar [18]. Application of the random forest seems to be an interesting alternative to the most often used neural networks.

An array of seven tin oxide-based gas sensors (TGS815, TGS821, TGS822, TGS825, TGS824, TGS842 and TGS822 modified by using an additional resistive potentiometer circuit) from Figaro Engineering Inc., mounted into an optimized test chamber has been applied in the computerized measurement system [2] (shape: cylinder, response time of sensors: 120 s, volume 0.2l of the test chamber was adjusted to the flow rate and time response). The carrier gas (synthetic air) flows through the chamber in controlled temperature conditions. The capacity of the measurement chamber, the carrier flow, the temperature and the size of gasoline sample are kept constant during the measurements. The signals are acquired by using an ADAM-4017 type 8-Channel Analog Input Module Rev.D1 and a serial communication interface with a PC computer.

The diagnostic features have been extracted from the averaged temporal series of sensor resistances $R(j)$, one for each j-th sensor ($j = 1, 2, \ldots, 7$) of the array. They are defined as relative variations $r(j)$ of each sensor resistance:

$$r(j) = \frac{R(j) - R_0(j)}{R_0(j)}, \tag{1}$$

where $R_0(j)$ represents the baseline resistance of j-th sensor measured in the synthetic air atmosphere.

The measurement system parameters were as follows: a carrier flow 0.2 l/min, a size of the gasoline sample 100 ml, a capacity of the sample chamber 200 ml, a gasoline temperature 25ºC. The sampling rate of sensor resistance was 30 times per minute. The baseline resistance $R_0(j)$ was registered at a stabilized temperature of 25ºC in a synthetic air. Its value was calculated by averaging 36 samples of the measured values within 72 s. A washing interval in the measurement was 10 min. The diagnostic feature vector **x** used in signal pattern recognition was composed of seven relative sensor signals described by (1) and was given in the form $\boldsymbol{x} = [r(1), r(2), ..., r(7)]^T$.

3. Data base

The experiments have been performed using pure extracted gasoline, characterized by the following physical and chemical properties: density – 0.665–0.700 g/cm3, final boiling point – 90ºC, relative content of aromatic hydrocarbons – 0.0005% [g/g]. The gasoline has been enriched by different supplements. They included ethanol, *ethyl tert-butyl eter (*ETBE), *methyl tert-butyl eter* (MTBE) and benzene, all of various concentrations in the blend. These components have been applied since they are most often used in petrol industry. Different types of blends, representing the classes under recognition, have been prepared [2]. The first four classes were formed by the extracted gasoline and ethanol concentrations of 5%, 10%, 15% and 20% of the volume. These concentrations have been chosen to reflect the recommended or accepted levels of bio-components in different countries (Brasil – 20%, USA – from 10% to 15%, Poland – 5%). The next four classes were created by adding MTBE and ETBE to the extracted gasoline in the proportion: MTBE (3%) and ETBE (97%). The last four blend families were created by adding benzene as a supplement. Benzene has a high octane number and thanks to this it is an important component added to the gasoline.

Four different blends of 5%, 10%, 15% and 20% concentrations for all mentioned supplements have been created in this way. The detailed description of classes is given in Table 1. Each class is represented by 72 carefully measured samples at a temperature of 25ºC, prepared from two different deliveries. The total number of samples is 864 and their acquisition has been done on the same day. The data analysed here were taken from [2], where only the original, not disturbed measurements, have been considered. They form the nominal set of data.

An additional set of sensor samples has been registered at an increased temperature (by heating the room space) of around 32ºC to observe the influence of temperature changes. These measurements have been done on another day and lasted a few hours. The environmental temperature was slightly changed from 31ºC to 34ºC in a random way. These samples have been normalized using the previously measured baseline resistance of sensors estimated at a basic temperature of 25ºC.

Table 1. The classes of gasoline blend used in the experiments.

Class	Type of additive
1	Ethanol additive of 5% volume
2	Ethanol additive of 10% volume
3	Ethanol additive of 15% volume
4	Ethanol additive of 20% volume
5	Additive of 5% volume (MTBE 3% and ETBE 97%)
6	Additive of 10% volume (MTBE 3% and ETBE 97%)
7	Additive of 15% volume (MTBE 3% and ETBE 97%)
8	Additive of 20% volume (MTBE 3% and ETBE 97%)
9	Benzene additive of 5% volume
10	Benzene additive of 10% volume
11	Benzene additive of 15% volume
12	Benzene additive of 20% volume

Figure 1 presents the influence of the environmental temperature on the measured signals from all sensors [19]. The solid line represents the basic results of measurements at a temperature of 25ºC and the dashed line the measurement results made at an increased temperature of around 32ºC.

An important task of the paper is to study the statistical behaviour of the electronic nose system in the existence of the disturbing noise. The zero mean Gaussian distribution white noise of different variance has been assumed to represent the possible measurements made in different environmental conditions. Different noise levels have been used in the experiments. The *signal-to-noise* (SNR) ratio varied from 60 dB to 0 dB. The SNR was defined in a standard way as the logarithm of the ratio of autocorrelation R_{ss} of signal and autocorrelation R_{nn} of noise, $SNR = 10\log\frac{R_{ss}}{R_{nn}}$. The SNR measured totally for all sensor signals has been assumed.

The classification experiments at this point aimed to check the robustness of the nose systems to the possible distortion in the measurement process. The strategy of cross-validation of data has been used in the experiments. The whole data set was split randomly into two equal parts. One part was used in the learning and the second – only in the testing mode. Two types of classification systems have been used: the random forest and the support vector machine. The random splits of data have been repeated 10 times changing the contents of learning and testing subsets. A percentage of testing error is estimated as the mean of the errors committed by the system in all runs.

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

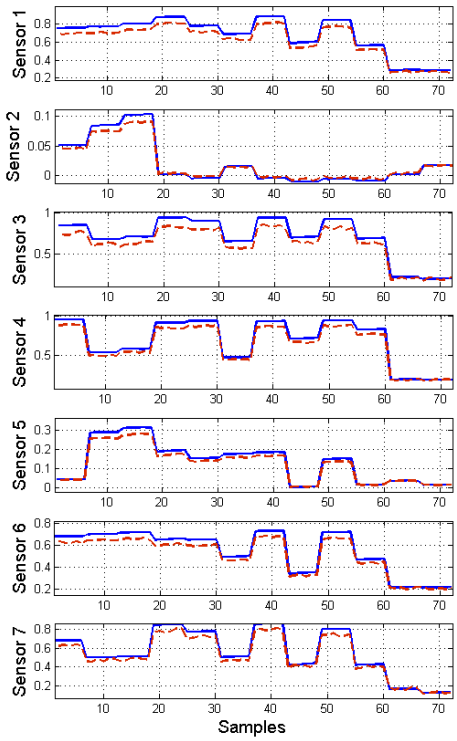

Fig. 1. The graphical effect of changing the environmental temperature for application of the same baseline resistance in both cases. The sensor signals represent normalized (dimensionless) values.

4. Self-organizing approach to mining the electronic nose data

4.1. Algorithms of clustering

Clustering of data consists in self-organizing division of n observations in an N-dimensional space into K subsets (clusters) represented by their centers $\mathbf{c}_i$, while providing the minimum total distance between the data vectors $\mathbf{x}_j$ and their winning centers $\mathbf{c}_i$ [20]. The process can be either unsupervised (without reference to a class membership of data) or supervised (a known class membership of data). In the case of unsupervised analysis the mathematical problem is described by:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in S_i} \left\| \boldsymbol{x}_j - \boldsymbol{c}_i \right\|^2. \tag{2}$$

The most popular solution is based on the competition technique. In the off-line implementation it is called K-means [20], whereas in the on-line mode its modification is known as *Conscience Winner Takes All* (CWTA). In this paper the on-line implementation is used, since it is less susceptible to the so called dead centers. Moreover, the on-line implementation leads usually to a smaller value of quantization error described by (2). In this technique each vector $\mathbf{x}_j$ is associated with its nearest center, which is subject to adaptation according to the relations [20]:

$$\boldsymbol{c}_i(k+1) = \boldsymbol{c}_i(k) + \eta\left[\boldsymbol{x}_j - \boldsymbol{c}_i(k)\right], \tag{3}$$

where η is a learning coefficient (usually $\eta < 1$), reduced to zero from iteration to iteration. To avoid the dead centers, the distance between the vector **x** and its winning center takes into account the past activity of the center by multiplying the real distance $d_r(\mathbf{x}_j,\mathbf{c}_i)$ by the number N_i of winnings of i-th center in the previous competitions, $d(\mathbf{x}_j,\mathbf{c}_i) \leftarrow N_i d_r(\mathbf{x}_j,\mathbf{c}_i)$ for $i = 1,2,\ldots,K$. At the beginning of the process $N_i = 1$ is assumed for all centers. This modification of distance calculation is applied only in the first cycle of adaptation.

After the segmentation stage the clusters are created, each containing vectors **x** of a high similarity. From a classification point of view the best case is when the data belonging to one class occupy similar locations in the space and are represented by a single cluster.

The quality of unsupervised clustering might be assessed by using different measures [20], from which the most often used are its purity and precision. The purity measures the extent to which a cluster contains objects of the same class. The purity of i-th cluster is defined as $p_i = \max_j p_{ij}$, where p_{ij} represents the percentage of representatives of j-th class in ith cluster. The overall purity of clustering n data points into K clusters is $p = \sum_{i=1}^{K} \frac{n_i}{n} p_i$, where n_i is the population of i-th cluster. The class precision is defined as the fraction of a cluster that consists of samples of the specified class. The precision of cluster i with respect to class j is $prec(i,j) = p_{ij}$.

In the case of supervised clusterization the samples belonging to one class are grouped into a separate cluster of the center being the average of data of this class. The quality of supervised clustering is usually assessed on the basis of cluster cohesion (compactness) and separation. The cluster cohesion defines the similarity of data points **x** to their cluster center **c**. It is usually expressed as the averaged distance between the data points and their representative center. On the other hand, the separation between two clusters is characterized by the distance between their centers.

4.2. Results of clustering experiments

The first experiments have been directed to finding relations between the class and the cluster membership of data. The CWTA algorithm was applied. The number of clusters was equal to the number of classes (12 in the experiments). Table 2 presents the percentages of classes belonging to particular clusters for the original data obtained at a nominal temperature of 25ºC. It is seen that all clusters are very good representatives of the classes and their class precision is 100%.

Table 2. The percentage of class membership in particular clusters obtained in CWTA self-organization for the original samples measured at a temperature of 25ºC.

CLUSTER \ CLASS	1	2	3	4	5	6	7	8	9	10	11	12
1	100											
2		100										
3			100									
4				100								
5					100							
6						100						
7							100					
8								100				
9									100			
10										100		
11											100	
12												100

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

The situation has changed significantly after the measurements performed at an increased temperature of around 32ºC. The measured data have been normalized by using the same value of the previous baseline corresponding to the temperature 25ºC. Table 3 presents the class distribution of such data. Here, we can observe some mixture of classes belonging to the same clusters. Only four clusters (2, 3, 5 and 12) contain the samples of a single class. The class precision of some clusters is low.

This distribution of classes in the presence of temperature changes is an evidence of the influence of the environmental temperature on the signal patterns representing different classes of data.

Table 3. The percentage of class membership in particular clusters obtained in CWTA self-organization after changing the environmental temperature.

CLUSTER \ CLASS	1	2	3	4	5	6	7	8	9	10	11	12
1	94.4								5.6			
2		100.0										
3			100									
4				88.9		33.3						
5					100							
6				11.1		66.7						
7							80.6	11.1		5.6		
8							11.1	80.6		5.6		
9	5.6								94.4			
10								8.3		88.8	8.3	
11							8.3				91.7	
12												100

An important issue in class recognition is the relative distribution of classes, especially distances between class centers. The larger this distance and the smaller standard deviation of data the easier the recognition task. This is especially important in an electronic nose, since the semiconductor sensor signals are vulnerable to noise caused by the change of environmental parameters. Table 4 shows these distances for all class centers (the mean of data belonging to the successive classes) for the samples measured at a nominal temperature. Large changes of these values are observed for different classes.

Similar calculations performed for the data registered at an increased temperature have shown that the centres have not changed their locations in a significant way. The maximum changes of centre locations for the perturbed data did not exceed a relative value of 5%.

Table 4. The distances between class centers for the nominal samples of data.

Class	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0.54	0.52	0.27	0.16	0.60	0.28	0.55	0.20	0.44	1.29	1.32
2	0.54	0	0.078	0.59	0.55	0.27	0.61	0.54	0.58	0.52	0.99	1.02
3	0.52	0.07	0	0.55	0.52	0.32	0.57	0.57	0.54	0.54	1.05	1.08
4	0.27	0.59	0.55	0	0.66	0.13	0.08	0.70	0.10	0.63	1.42	1.46
5	0.16	0.55	0.52	0.66	0	0.62	0.16	0.61	0.09	0.52	1.35	1.38
6	0.60	0.27	0.32	0.13	0.62	0	0.71	0.36	0.67	0.42	0.79	0.82
7	0.28	0.61	0.57	0.08	0.16	0.71	0	0.08	0.75	0.67	0.47	1.5
8	0.55	0.54	0.57	0.70	0.61	0.36	0.08	0	0.69	0.17	0.80	0.83
9	0.20	0.58	0.54	0.10	0.09	0.67	0.079	0.69	0	0.60	1.42	1.45
10	0.44	0.52	0.54	0.63	0.52	0.42	0.67	0.17	0.60	0	0.30	0.93
11	1.29	0.99	1.05	1.42	1.35	0.79	0.47	0.80	1.42	0.30	0	0.09
12	1.32	1.02	1.08	1.46	1.38	0.82	1.5	0.83	1.45	0.93	0.09	0

The distances of data samples to their winning center is another important factor that should be taken into account in the analysis. The average distance of samples to their winning center and the standard deviation are significantly dependent on the temperature. Table 5 presents these values for the data registered at nominal and increased temperatures. The third column shows the comparative results for the nominal data distorted by random noise of normal distribution at SNR = 12 dB.

Table 5. The average distances of samples to their centers and standard deviations for the data registered at 25ºC, around 32ºC and the data distorted by random noise of normal distribution at SNR = 12 dB.

Class	Data registered at temperature 25ºC	Data registered at temperature ~32ºC	Nominal data distorted by random noise of SNR = 12 dB
1	0.005 ± 0.0036	0.014 ± 0.0136	0.054 ± 0.0321
2	0.005 ± 0.0045	0.024 ± 0.0229	0.025 ± 0.0230
3	0.003 ± 0.0028	0.015 ± 0.0148	0.016± 0.0134
4	0.020 ± 0.0143	0.024 ± 0.0162	0.048 ± 0.0351
5	0.006 ± 0.0046	0.018 ± 0.0193	0.068 ± 0.0433
6	0.012 ± 0.0074	0.015 ± 0.0136	0.040 ± 0.0305
7	0.005 ± 0.0044	0.026 ± 0.0272	0.033 ± 0.0190
8	0.025 ± 0.0168	0.028 ± 0.0256	0.044 ± 0.0374
9	0.003 ± 0.0023	0.030 ± 0.0268	0.034 ± 0.0164
10	0.008 ± 0.0057	0.023 ± 0.0313	0.039 ± 0.0356
11	0.005 ± 0.0035	0.017 ± 0.0125	0.052 ± 0.0377
12	0.003 ± 0.0022	0.019 ± 0.0184	0.050 ± 0.0258

The results show that the distorted data are characterized by significantly larger average distances from their winning centers (higher dispersion). This increase depends on the amount of noise and – in the case of SNR = 12dB – the largest observed relative increase is almost 17 (for the 12th class of data). It means that in this case some distorted samples are closer to the neighbouring centers than to their own, which is equivalent to a misclassification.

To obtain a graphical presentation of the distribution of classes we have mapped the 7- dimensional data onto two dimensions by applying the *principal component analysis* (PCA) of the measured samples. The PCA [20] is a linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$, mapping the N- dimensional original vector $\mathbf{x}$ onto a K-dimensional output vector $\mathbf{y}$, of $K < N$ (in our case $K = 2$). The transformed vectors $\mathbf{y}$ preserve the most important features of the original information. The PCA matrix $\mathbf{W}$ is composed of the most important eigenvectors of a covariance matrix $\mathbf{R}_{\mathbf{xx}}$ built for the set of input vectors $\mathbf{x}_i$. The 2-dimensional coordinate system in this analysis is created by the first two most important principal components $PC_1 = y_1$ and $PC_2 = y_2$.

Figure 2a shows the results of PCA analysis of the originally (non-disturbed) measured samples. The samples belonging to different classes have been denoted by the letter C and the successive number of class, *i.e.* C1, C2, *etc*. The presented distribution of these mapped samples is an evidence that the points belonging to particular classes of gasoline blends create separated compact clusters. Twelve distinct gasoline clusters, each composed of samples belonging to the same class, are easily recognized. The clusters are well separated from each other and characterized by a relatively small dispersion. This is a very good prognostic for accurate recognition and separation of all classes.

Adding noise to the measured samples introduces fuzziness in the data distribution (Fig. 2b). The clusters representing different classes occupy now a larger space and the samples belonging to the closest clusters interlace each other, making the class recognition problem much more difficult. For example, six classes on the right side of the figure form now a completely mixed

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

environment of data. The higher the level of distortion the more fuzzy character of the cluster distribution is observed. Therefore, applying the clustering in the process of class recognition based on the purity of clusters in a noisy environment is inefficient and leads to a very large degree of misclassification.

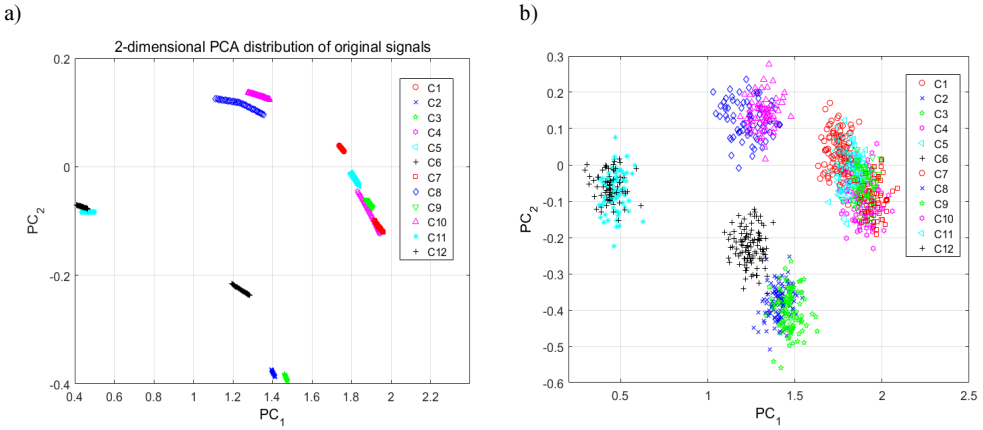

Fig. 2. 2-dimensional PCA plots of the data samples: the original measured data (a); the data corrupted by noise with SNR = 12 dB (b).

5. Supervised approach to data mining

5.1. Supervised classifiers

The supervised approach includes application of two types of classifiers: the *random forest* of decision tree (RF) and the *support vector machine* (SVM).

The random forest proposed by Breiman [11] is an ensemble of many multivariate decision trees. It constructs decision trees in the training time and outputs a class, that is the mode of classes pointed by individual trees (majority voting). The learning data (often 2/3 of the whole data set) are selected randomly for each tree.

A small group of input variables to split is selected at random in each node of a decision tree. The group size m is fixed. The linear combinations of m randomly selected variables in each node are generated, and then a search is made over them for the best split. The predicted variables that provide the best split according to a predefined objective function are used to do a binary split on that node. In the next nodes, other m variables are chosen at random from all predicted variables and previous operations are repeated. After generating a large number of trees they vote for the most popular class.

It was proved [11] that random split selection of data increases the recognizing system immunity to noise contained in the measurement data. This is a very important property of an electronic nose, since the sensor signals may experience some unpredictable variations.

The application of RF enables to assess the significance of individual sensors for the performance quality of an electronic nose. The impact of a particular sensor signal is estimated by taking into account its influence on the classification results, in particular, how inclusion of this signal is important for getting a higher accuracy of class recognition [11].

Generally, the importance of input attribute in RF is measured by an increase of prediction error for the validation data if the values of this attribute are permuted among the testing data. The out-of-bag prediction error is computed on this perturbed data set and compared with the error before perturbation. The higher this increase, the more important is the input attribute.

This measure is estimated for every tree, then averaged over the ensemble and divided by the standard deviation over the entire ensemble [11, 19]. In this way the input attributes (sensor signals) are ordered according to their statistical impact on the classification accuracy.

The *Support Vector Machine* (SVM) is a feedforward network of one hidden layer (the kernel function layer) known for its good generalization ability [12]. In the learning phase it constructs a hyperplane in a high-dimensional space, separating the learning vectors into two classes of the destination values either $d_i = 1$ (one class) or $d_i = -1$ (the opposite class), with the maximal separation margin (the largest distance of the nearest training data points of the opposite classes). The SVM model represents the original data as points mapped in the space in such a way that examples of different categories are separated by clear gaps that are as wide as possible. The width of separation margin formed in the learning stage depends on a regularization constant C, which should be properly adjusted by the user. Thanks to such a learning strategy the network is resistant to noise contaminating the input data.

A great advantage of SVM is unique formulation of the learning problem leading to the quadratic programming with linear constraints, which is easy to solve. The SVM of Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|)$, treated as the most universal one, has been used in this application. The hyper-parameters γ of the Gaussian function and the regularization constant C have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one in the validation data sets. The optimal values of these parameters found in the preliminary experiments were as follows: $\gamma = 1$ and $C = 1000$. They have been found by applying a trial-and-error process using the validation data (a third of the learning data volume). To deal with the problem of many classes we have applied a one-against-one strategy [12]. In this method many 2-class SVM classifiers are used for all combinations of two classes. The final classification decision is based on the majority voting principle.

5.2. Supervised classification results

The recognition ability of the classes of blends was checked by applying two systems of supervised classification: the support vector machine and the random forest of decision trees. The SVM of Gaussian kernel and hyper parameters $\gamma = 1$ and $C = 1000$ were applied in all 10 validation runs. On the other hand, the RF of 50 trees was constructed using 3 input variables selected at random in each node to split. These meta-parameter values of RF have been adjusted after some introductory experiments using the trial-and-error approach, in which different values of trees and node variables have been tried. The choice providing the best results on the validation data (a third of the learning data volume) has been accepted in further experiments. The RF experiments with random selection of learning and testing data have been also repeated 10 times and their results averaged.

Application of both classifiers to the recognition of original samples of blends registered at a temperature of 25°C has resulted in 100% accuracy of class recognition. These excellent results are in accordance with the class-cluster distribution presented in Table 2 and also with the PCA results of the original data presented in Fig. 2a.

However, introducing noise to the testing samples while training the classifiers on an undisturbed data set, has resulted in decreasing this accuracy. This reduction was dependent on the actual SNR value, the type of applied classification system and the number of applied sensors. Generally, the higher the noise, the larger degree of misclassification.

An important issue is the impact of individual sensors on the class recognition results. In solving this problem the random forest ability was used. A measure of the sensor importance in this method is defined as a relative increase of error after perturbation of its value compared with the error before this perturbation The more important sensor corresponds to a larger

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

relative increase of this error. Fig. 3 presents the result of application of random forest to the measure of class discrimination ability of successive sensors. The results show that all sensors contribute positively to the classification results and their influence is at a similar, although not equal, level.

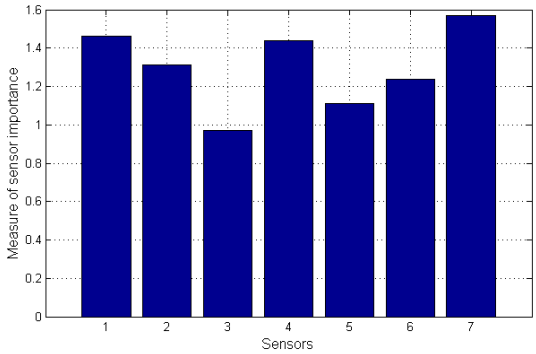

Fig. 3. The importance of sensors in class recognition according to the random forest measure.

The next experiments have been performed with a gradually reduced number of sensors. In the sensor elimination process the embedded property of random forest to assessing the importance of individual sensor signals in the classification process has been used. The results of this estimation are presented in Fig. 3. The least important in this phase of classification was the third sensor. Therefore, in the next runs of calculations this sensor signal was eliminated from the input vector as the first one. Next, the same procedure of estimation of importance of the remaining six sensors was repeated which to elimination of the next least important sensor (the importance of sensors is changing from run to run and depends on a set of remaining sensors).

The statistical results concerning the recognition error of the blend classes at different levels of noise and different quantities of applied sensors are presented in Table 6. The classifiers were trained on the undisturbed nominal data subsets and tested on the other subsets disturbed by artificial noise of different levels. The results refer to application of the SVM and RF classification systems to three different matrices of sensors, *i.e*. containing 7, 6 and 5 sensors. The table presents the mean values and *standard deviations* (std) of class recognition obtained in 10 runs of the classification process. Each run was associated with random noise of a specified SNR value.

There is a visible correlation between the number of applied sensors and robustness of the system to the noise level. The elimination of sensors leads to an increase of the recognition error. A type of classifier is also very important. The results show that SVM is much more resistant to noise than the random forest. This is well seen for high levels of noise. For example, at SNR = 20 dB and application of 7 sensors the mean error of class recognition for SVM was equal to 5.12% and for RF it increased to 19.87%. In the case of six sensors the respective values were 5.80% (SVM) and 21.14% (RF). An advantage of SVM over RF follows from the margin of separation built in the learning process of SVM, which is not the case for RF.

Figure 4 shows a plot of the mean values of relative recognition error versus SNR values for both classifiers and different numbers of sensors applied. These results have been obtained in 10 runs of the classification procedure. They confirm the superiority of SVM performance in the presence of noise.

Table 6. The mean errors of recognition of gasoline blends in application of RF and SVM for different levels of noise.

SNR [dB]	Mean error and std for 7 sensors [%]		Mean error and std for 6 sensors [%]		Mean error and std for 5 sensors [%]	
	SVM	RF	SVM	RF	SVM	RF
60	0	0	0	0.20 ± 0.12	0	0.30 ± 0.18
45	0	0.26 ± 0.19	0	0.32 ± 0.30	0	0.82 ± 0.25
52	0	0.80 ± 0.51	0	0.90 ± 0.53	0	1.01 ± 0.68
32	0	6.12 ± 1.62	0	8.71 ± 2.50	0	9.95 ± 1.23
25	0.37 ± 0.32	13.23 ± 3.4	0.39 ± 0.34	17.5 ± 1.59	0.56 ± 0.42	21.45 ± 3.63
20	5.12 ± 1.34	19.87 ± 3.26	5.80 ± 1.68	21.14 ± 2.84	6.9 ± 0.98	26.34 ± 3.82
15	10.87 ± 2.34	38.47 ± 3.78	19.68 ± 2.08	40.49 ± 3.92	22.78 ± 2.07	43.87 ± 4.03
10	27.21 ± 3.24	42.09 ±2.73	28.43 ± 4.02	44.34 ± 2.13	31.67 ± 4.62	47.58 ± 2.24
7.5	35.35 ± 2.41	53.24 ± 3.96	36.78 ± 2.64	55.47 ± 4.81	39.68 ± 2.34	60.56 ± 3.15
6	41.24 ± 3.24	62.89 ± 3.68	44.21 ± 3.91	65.39 ± 3.58	45.23 ± 3.17	68.12 ± 4.25
0	60.34 ± 2.95	80.25 ± 2.73	62.80 ± 1.98	83.73 ± 1.99	64.52 ± 3.02	85.89 ± 3.27

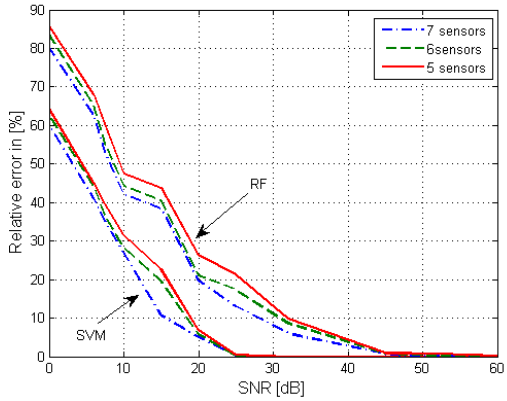

Fig. 4. The relative error of class recognition by RF and SVM for different levels of noise corrupting the measured data and different numbers of sensors.

6. Wavelet de-noising sensor signals

The nose measurements made in various environmental conditions produce non-deterministic sensor signals in different frequency domains with respect to the basic measurement, at which the baseline was estimated. Therefore, a proper transformation of sensor signals from the time to the frequency domain and a careful removal of noisy components can filter the sensor signals. The procedure presented in the paper was implemented by using the discrete wavelet transform.

6.1. Principle of wavelet de-noising

The *discrete wavelet transform* (DWT) is a linear transformation with a special property of simultaneous localization in time and frequency. It decomposes a given discrete signal series into a set of specially defined basic functions of different frequencies, shifted mutually and called wavelets [21, 22].

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

The aim of DWT is to decompose an analyzed signal $x(t)$ into a finite summation of wavelets of different scales (levels) and shifts according to the expansion:

$$x(t) = \sum_j \sum_k c_{jk} \psi(2^j t - k), \tag{4}$$

where c_{jk} is a new set of coefficients and $\psi(2^j t - k)$ is the wavelet of j-th level (scale) shifted by k samples. The set of wavelets of different scales and shifts can be generated from a single prototype mother wavelet, by dilations and shifts. In practice, most often used are the orthogonal or bi-orthogonal wavelet functions, forming an orthogonal or bi-orthogonal base [19, 21].

Denote a discrete form of the original signal vector by $\mathbf{x}$ and by $A_j\mathbf{x}$ an operation that computes the approximation of $\mathbf{x}$ with a resolution 2^j. Let $D_j\mathbf{x}$ denotes the detailed signal, $D_j\mathbf{x} = A_{j+1}\mathbf{x} - A_j\mathbf{x}$, defined as the difference of signal approximations for two neighboring resolutions. Both operations $A_j\mathbf{x}$ and $D_j\mathbf{x}$ can be interpreted as the convolution of the signal of previous resolution and the finite impulse response of the quadrature mirror filters: the high-pass one ($\widetilde{G}$) with coefficients $\tilde{g}$ and the low-pass one ($\widetilde{H}$) with coefficients $\tilde{h}$:

$$A_j \boldsymbol{x} = \sum_{k=-\infty}^{\infty} \tilde{h}(2n-k) A_{j+1} \boldsymbol{x}(2n), \tag{5}$$

$$D_j \boldsymbol{x} = \sum_{k=-\infty}^{\infty} \tilde{g}(2n-k) A_{j+1} \boldsymbol{x}(2n). \tag{6}$$

These two operations performed at different levels, from $j = 1$ to J, deliver the decomposition coefficients for different scales and resolutions of the original signal $\mathbf{x}$. The most often used discrete wavelet analysis uses the Mallat pyramid algorithm [22].

The result of such transformation is a set of coefficients representing the detailed signals $D_j\mathbf{x}$ at different levels j ($j = 1, 2,.., J$) and the residual signal $A_J\mathbf{x}$ at the level J. All of them are of different resolution, characteristic for the applied level. The $D_j\mathbf{x}$ can be interpreted as the high frequency details, distinguishing the approximations of the signal for two neighboring levels of resolution. The signal $A_J\mathbf{x}$ represents a coarse approximation of the vector $\mathbf{x}$.

Transformation of the detailed signals $D_j\mathbf{x}$ ($j = 1, 2,..., J$) and the coarse approximation signals $A_J\mathbf{x}$ into the original resolution is possible using special filters G and H associated with the analysis of filters $\widetilde{G}$ and $\widetilde{H}$ by the quadrature and reflection relationships [21, 22]. This is done by the reverse Mallat pyramid algorithm. The original signal $\mathbf{x}(n)$ at each time instant n is reconstructed by simply adding appropriate wavelet coefficients and the coarse approximation, both transformed to the same original resolution. At J-th level of DWT we have:

$$x(n) = D_1(n) + D_2(n) + ... + D_J(n) + A_J(n). \tag{7}$$

Figure 5 presents the results of 4-level DWT of the sensor data in measurement of petrol with bio-additives after adding noise [19].

The Haar wavelets have been applied in the decomposition. The first four levels of wavelet coefficients represent the detailed coefficients from D_1 to D_4, whereas the next one – denoted by A_4 – the coarse approximation at the 4th level. All are presented in the original resolution. We observe a substantial difference of variability of signals at different levels. The first level detail coefficients D_1 represent the highest variability of signal, which is usually associated with the high frequency noise.

Application of the wavelet transformation in sensor technology is not new. It was used *e.g.* for extraction of features in porous silicon chemical sensors [23]. Our idea is to apply this transformation to reduction of noise. This is done by cutting the lowest value detailed coefficients of wavelet decomposition and reconstruct the sensor signals deprived of them. The cut coefficients are treated as the noise components. Thanks to this we obtain reduction of noise contaminating the measured signals.

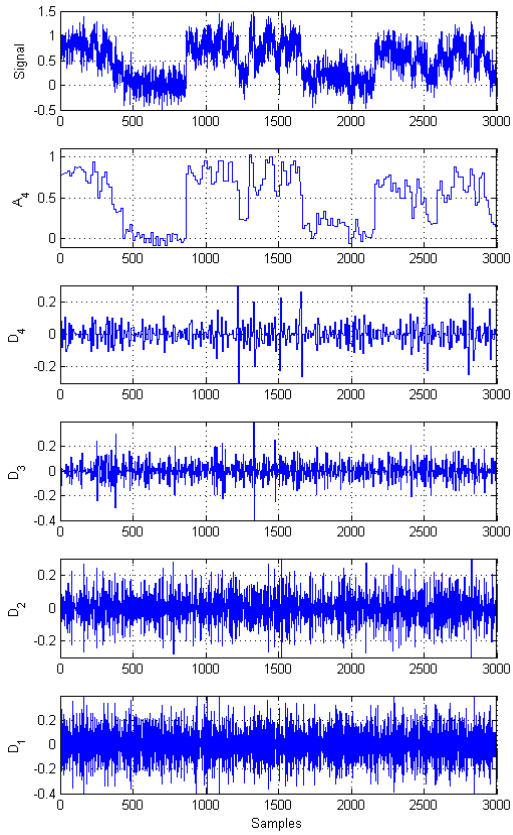

Fig. 5. DWT of the measured sensor signals. D_1 to D_4 represent the detailed coefficients and A_4 the coarse approximation of signal at the 4th level. The signals are represented by the normalized (dimensionless) values.

6.2. The results of de-noising using DWT

The next experiments have been performed using DWT for de-noising the sensor signals artificially distorted by the white noise with a normal distribution and a different variance. The aim is to reduce the noise in the data and in this way to increase the probability of proper pattern recognition. The de-noising process is performed by decomposing the noisy sensor signals into a few decomposition levels and then reconstructing them by eliminating the least important details. The four-level decomposition using Haar wavelets and soft thresholding of fixed values in each detail coefficient have been found to be the best. The results of such de-noising are presented in Fig. 6.

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

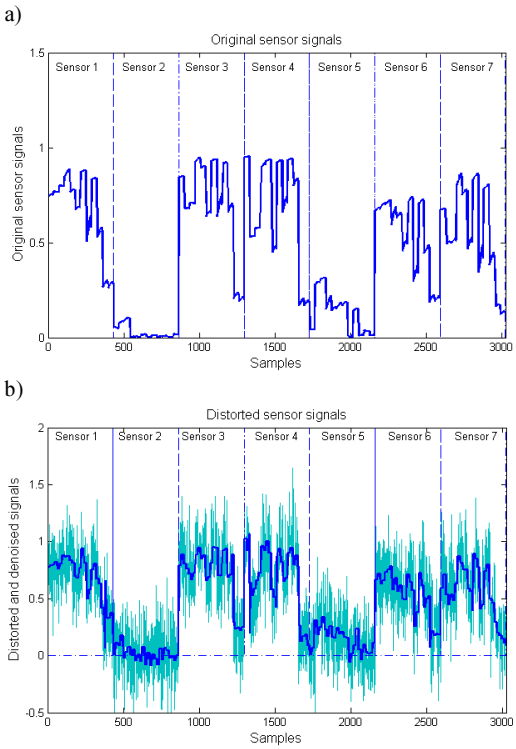

Fig. 6. An illustration of the de-noising process of sensor signals: the original samples of undistorted sensor signals (a); the de-noised signal (blue color) with the noisy signal at SNR = 0 dB at the background (green color) (b). The signals are represented as the normalized (dimensionless) values.

The upper Fig. 6a presents the original samples of undistorted sensor signals and the bottom one (Fig. 7b) the de-noised samples with the distorted signals at SNR = 0 dB at the background. A significant decrease of noise is visible. Fig. 7 illustrates the effect of signal de-noising by presenting the relative difference between the original sensor signals $x(n)$ and the distorted signals $x_{noise}(n)$ and between the original signals and the signals after de-noising $x_{denoise}(n)$. This difference is expressed in the form of relative coefficients α_{noise} and $\alpha_{denoise}$:

$$\alpha_{noise} = \frac{\|x(n) - x_{noise}(n)\|}{\|x(n)\|}, \tag{8}$$

$$\alpha_{denoise} = \frac{\|x(n) - x_{denoise}(n)\|}{\|x(n)\|}. \tag{9}$$

The effect of de-noising is visible, especially at high values of noise (SNR close to zero). The reduction ratio of noise contents in the signal in such a case well exceeds 2.

This effect is also very well seen in the 2-dimensional coordinate system formed by two most important PCA components of the sensor signals. This is illustrated in Fig. 8.

Figure 8a presents the distribution of noisy samples of the sensor signals at SNR = 0 dB and Fig. 8b their distribution after the de-noising process. In the first case (Fig. 8a) the samples belonging to different classes are completely mixed, while after de-noising (Fig. 8b) the representatives of individual classes are grouped together and the classes are reasonably well separated from each other.

The de-noised sensor signals have been used as the input attributes to the SVM classifier in the process of class recognition. The classifier was trained on the original (undistorted) part

of data and then tested on a separate part of distorted data before and after de-noising. The average results of 10 runs in such organized classification process at different noise levels are presented in Table 7.

The table presents the mean percentage errors and their standard deviations obtained in 10 runs of the classification process for the noisy data before and after their de-noising. The results are given for 7 sensors and after reducing their numbers to 6 and 5, preserving the most important sensors. A significant reduction of classification errors can be observed for each level of noise.

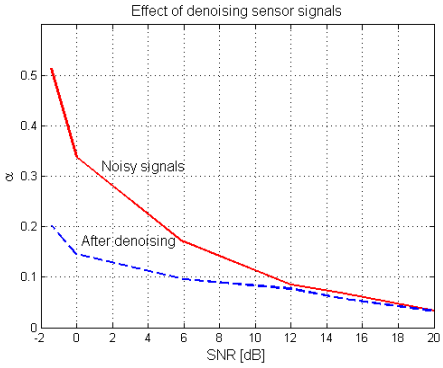

Fig. 7. The influence of the de-noising process on the reduction of noise contents in the signals.

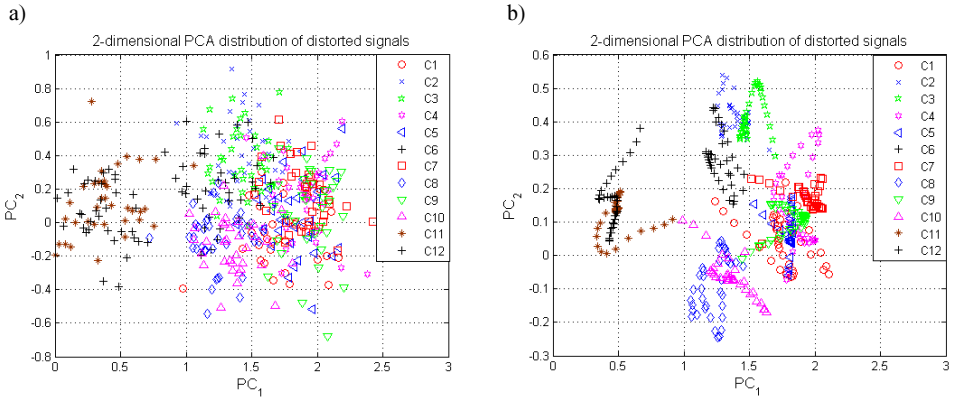

Fig. 8. The distribution of sensor signal samples of 12 classes mapped onto the 2-dimensional coordinate system formed by two most important principal components PC_1 and PC_2: signals distorted by noise of SNR = 0 dB (a); signals after the DWT de-noising process (b).

Table 7. The SVM classification results of de-noising.

SNR [dB]	Mean error and std for 7 sensors [%]		Mean error and std for 6 sensors [%]		Mean error and std for 5 sensors [%]	
	noisy signals	after de-noising	noisy signals	after de-noising	noisy signals	after de-noising
20	5.12 ± 1.34	2.7 ± 0.63	5.80 ± 1.68	3.5 ± 0.71	6.9 ± 1.17	4.7 ± 0.98
15	10.87 ± 2.34	4.4 ± 0.91	19.68 ± 2.08	6.3 ± 1.03	22.78 ± 2.07	10.9 ± 1.34
12	27.21 ± 3.24	12.3 ± 1.67	28.43 ± 4.02	12.9 ± 1.89	31.67 ± 4.62	14.7 ± 2.23
6	41.24 ± 3.24	18.1 ± 2.54	44.21 ± 3.91	19.2 ± 2.79	45.23 ± 3.17	26.5 ± 2.96
0	60.34 ± 2.95	35.1 ± 1.71	62.80 ± 1.98	35.9 ± 1.87	64.52 ± 3.02	36.3 ± 2.03

Metrol. Meas. Syst., Vol. 24 (2017), No. 1, pp. 27–44.

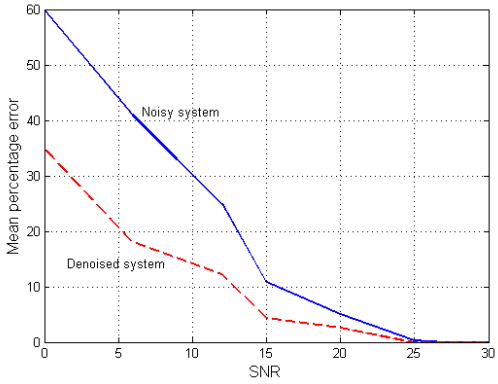

Fig. 9. The dependence of the mean recognition error on the SNR value for 7 noisy sensor signals before and after their de-noising.

Figure 9 shows this difference for 7 sensors in a graphical way. It presents the dependence of the mean percentage value of class recognition error on signal de-noising obtained for all 7 sensors obtained for the noisy sensor signals .

7. Conclusions

The paper presents the analysis of data distorted by noise, simulating a noisy environment of electronic nose measurement of the gasoline blends. The array of semiconductor sensors, forming the heart of an electronic nose, corresponds with a signal pattern characteristic for each gasoline blend type. The pattern recognition system working in the classification mode processes these signals and associates them with an appropriate class. The problem is that each nose measurement includes some distortion resulting from changing environment conditions and fluctuation of sensor signals. The examinations aimed in two directions: the unsupervised analysis based only on the measured sensor signals, and the supervised analysis, where the class represented by a signal pattern is known in the learning phase. It was proved that the noise corrupting the measurement has a significant influence on the locations of clusters, and also increases their dispersion, making the pattern recognition based on distances not efficient.

Therefore, the research was directed to finding the most efficient supervised classification systems. Basing on the actual experience in this field two most efficient classifiers have been chosen: the RF and SVM ones. The main advantage of RF is the fact that it never over fits. Injecting a right kind of randomness in each step of signal processing at a large number of grown trees improves the generalization ability of the system and makes RF a good solution to the classification problems. On the other hand, the SVM network is resistant to noise because of the maximized separation margin between two recognized classes, formed in the learning phase. Such a property makes this tool suitable for working in noisy environments.

Both classification systems have been tested on samples of the gasoline blends formed on the basis of a few bio-based additives injected in different concentrations. In the case of original (noiseless) data, 100% accuracy of class recognition irrespectively of the applied classifiers has been achieved. However, adding noise to the measured samples has decreased the accuracy. In this case SVM has shown its superiority over RF, because of its large insensitivity to noise due to the separation margins between different classes introduced in the learning phase.

The results of class recognition in the presence of noise have confirmed that an electronic nose built on the basis of SVM is a better solution to the pattern recognition of gasoline blends.

The last direction of research was to examine the possibility of de-noising sensor signals. The DWT has been applied in this process. By cutting the small (insignificant) detail

coefficients a great reduction of noise has been achieved. Such de-noised sensor signals applied as the input attributes to the classifiers have increased the accuracy of pattern recognition and the efficiency of final classification system.

References

[1] McCarrick, C.W., Ohmer, D.T., Gillil L.A., Edwards, P.A. (1996). Fuel identification by neural network analysis of the response of vapour-sensitive sensor arrays. *Analytical Chemistry*, 68, 4264–4269.

[2] Brudzewski, K., Osowski, S., *et al*.(2006). Classification of gasoline with supplement of bio-products by means of an electronic nose SVM neural network. *Sensors and Actuators B*, 113, 135–141.

[3] Di Natale, C., Martinelli, E., D'amico, A. (2005). Pre-processing and pattern recognition methods for artificial olfaction systems: a review. *Metrol. Meas. Syst*., 12(1), 3–26.

[4] Bielecki, Z., Janucki, J., *et al*. (2012). Sensors and systems for the detection of explosive devices – an overview. *Metrol. Meas. Syst.*, 19(1), 3–28.

[5] Boeker, P. (2014). On 'Electronic Nose' methodology. *Sensors and Actuators B*, 204, 2–17.

[6] Jha, S.K., Yadava, R.D. (2011). Denoising by singular value decomposition its application to electronic nose data processing. *IEEE Sensors Journal*, 11, 1, 35–44.

[7] Hassanpour, H. (2008). A time-frequency approach for noise reduction. *Digital Signal Processing*, 18, 728–738.

[8] Fonollosa, J., Fernández, L., *et al*. (2016). Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization. *Sensors and Actuators B*, 236, 1044–1053.

[9] Zuppa, M., Distante, C., Siciliano, P., Persaud, K.C. (2004). Drift counteraction with multiple self-organizing maps for an electronic nose. *Sensors and Actuators B*, 98, 305–317.

[10] Kalinowski, P., Jasiński, G., Jasiński, P. (2014). Stabilność odpowiedzi półprzewodnikowych czujników gazu w zmiennych warunkach środowiskowych: badania długoterminowe oraz korekcja dryftu. *Elektronika: konstrukcje, technologie, zastosowania*, 55(9), 119–121.

[11] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

[12] Schölkopf, B., Smola, A. (2002). *Learning with kernels*. Cambridge MA: MIT Press.

[13] Wiziack, N.K.L., Catini, A., Santonico, M., D'Amico, A., Paolesse, R., Paterno, L.G., Fonseca, F.J., Di Natale, C. (2009). A sensor array based on mass capacitance transducers for the detection of adulterated gasolines. *Sensors and Actuators B*, 140, 508–513.

[14] Osowski, S., Tran Hoai, L., Brudzewski, K. (2004). Neuro-fuzzy TSK network for calibration of semiconductor sensor array for gas measurements. *IEEE Trans. on Measurements Instrumentation*, 53, 330–637.

[15] Guney, S., Atasoy, A. (2012). Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm support vector machine in an electronic nose. *Sensors Actuators B*, 166–167, 721–725.

[16] Botre, B.A., Gharpure, D.C., Shaligram, A.D. (2010). Embedded electronic nose supporting software tool for its parameter optimization. *Sensors and Actuators B*, 146, 453–459.

[17] Pardo, M., Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. *Sensors and Actuators B*, 107, 730–737.

[18] Liu, M., Wang, M., Wang, J., Li, D. (2013). Comparison of random forest, support vector machine back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage Chinese vinegar. *Sensors and Actuators B*, 177, 970–980.

[19] Matlab user manual (2014). Natick, USA: MathWorks.

[20] Tan, P.N., Steinbach, M., Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education Inc.

[21] Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM, Philadelphia.

[22] Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–692.

[23] Murguía, J.S., Vergara, A., Vargas-Olmos, C., Wong, T.J., Fonollosa, J., Huerta, R. (2013). Two-dimensional wavelet transform feature extraction for porous silicon chemical sensors. *Analytica Chimica Acta*, 785, 1–15.