Piotr S. SZCZEPANIAK, Marcin KŁOSIŃSKI

Institute of Information Technology, Technical University of Łódź

Maximal margin classifiers applied to DGA-based diagnosis of power transformers

Abstract. *The paper addresses a modern approach to the problem of power transformer diagnosis. The method called support vector machines enables the creation of an expert system for oil transformer technical condition diagnosis. The system, which is based on real results of chromatography of gases dissolved in transformer oil (DGA), performs better than an internationally acknowledged standard – the IEC code.*

Streszczenie. *Przedstawiono nową metodę diagnostyki transformatora mocy bazująca na algorytmie "support vector machine". W systemie bada się chromatograficznie gazy rozpuszczone w oleju transformatorowym. (**Zastosowanie klasyfikatora do diagnozy gazów rozpuszczonych w oleju transformatora mocy**)*

Keywords: power transformers, DGA, IEC, classification, maximal margin classifiers, SVM
Słowa kluczowe: transformator mocy, analiza oleju, klasyfikator, SVM.

Introduction to DGA-based diagnosis

The analysis of gases dissolved (DGA) in oil is one of the basic methods applied to diagnosing technical condition of transformers. The internationally acknowledged rules for interpretation of concentration of gases are reflected in the IEC code [IEC(1979)]. However, basing on their own experience, different countries have worked out different transformer diagnosing methods. All those methods base on rigid mathematical dependencies.

Data used to evaluate a transformer are defined by three variables *(x,y,z)* as shown in (1).

(1) $$x = \frac{C_2H_2}{C_2H_4} \qquad y = \frac{CH_4}{H_2} \qquad z = \frac{C_2H_4}{C_2H_6}$$

where H_2, CH_4, C_2H_2, C_2H_4, and C_2H_6 denote the amount of hydrogen, methane, acetylene, ethylene, and ethane in the gas under examination (in ppm units – parts per one million), respectively. The meaning of these variables is the same as in the IEC code; they reflect the DGA results. Following the IEC code, it has been assumed that there are nine classes describing the state of the transformer:

- No fault
- Partial discharge of low energy
- Partial discharge of high energy
- Disruptive discharge of low energy
- Disruptive discharge of high energy
- Overheating below 150°C
- Overheating between 150oC and 300°C,
- Overheating between 300oC and 700°C
- Overheating over 700°C.

Consequently, every triple enables reasoning about the technical condition of the examined transformer.

The ranges of quotient values are coded with three integer numbers {0,1,2} (see Table 1.). According to the IEC code, there are nine classes (see Table 2.), which are properly defined from an engineering perspective with an additional class containing unrecognized cases.

Such expert knowledge can be represented in the form of logic rules:

(2) If *code*$\{d_x, d_y, d_z\}$, then *transformer's state*.

These are crisp logic rules, which are unambiguous and mutually exclusive.

Table 1. IEC codes

Transformer state \ code	$\frac{C_2H_2}{C_2H_4}$	$\frac{CH_4}{H_2}$	$\frac{C_2H_4}{C_2H_6}$
Without fault	0	0	0
Partial discharge of low energy	0	1	0
Partial discharge of high energy	1	1	0
Disruptive discharge of low energy	1 or 2	0	1 or 2 *
Disruptive discharge of high energy	1	0	2
Overheating below 150 °C	0	0	1
Overheating between 150 °C and 300 °C	0	2	0
Overheating between 300 °C and 700 °C	0	2	1
Overheating over 700 °C	0	2	2
Unrecognized fault	Other codes		

Table 2. Classification according to IEC code

Value \ quotient	$\frac{C_2H_2}{C_2H_4}$	$\frac{CH_4}{H_2}$	$\frac{C_2H_4}{C_2H_6}$
[0; 0,1)	0	1	0
[0,1; 1)	1	0	0
[1; 3)	1	2	1
Value > 3	2	2	2

Visually, the IEC coding system creates a separation in three-dimensional space by introducing a set of disjoint fragments (Fig.1). Of course, the biggest fragment is the one representing an unrecognized state.

To summarize, to each triple (x, y, z), or to each triple presented in Table 2 one can assign an additional value describing (in an encoded or linguistic way) the technical condition of a transformer. This value, which is figured out on the basis of the inspection of an unplugged transformer or an expert opinion, can be actually interpreted as a diagnosis of the object's condition.

The idea presented in this paper is as follows. Having a sufficient number of representative results for a series of transformers, one should be able to find regions related to the above nine classes. The examination of a new DGA result would be then straightforward. So far, the IEC, as well as every other national system of classification, has offered the arbitrary division of the space. In the presented proposition, the determination of regions related to the nine

*) Code „102" is usually interpreted as „full discharge of high energy", that is, the diagnosis „full discharge of low energy" is associated with one of the three codes: {101}, {201} or {202}.

classes of diagnosis depends on the real results of previous examinations performed on transformers working in real conditions. The regions are determined using the support vector machines.

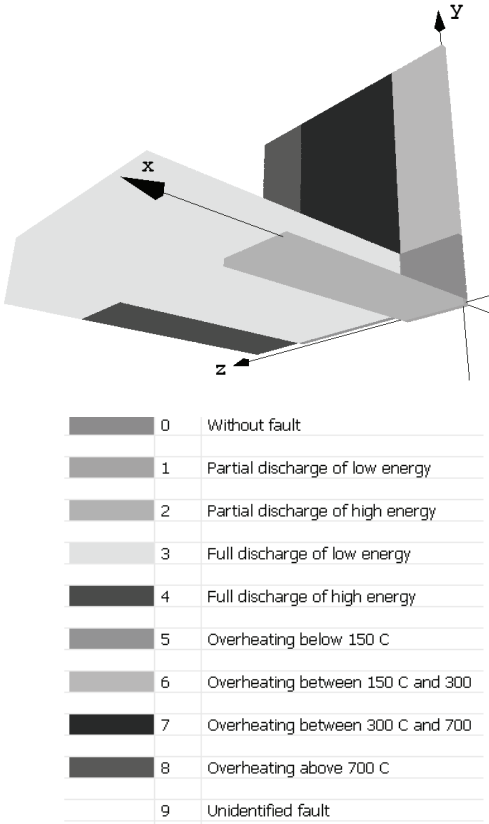

Fig.1. Visualization of IEC codes classification

Basics of Support Vector Machines

The presented classifiers are *Support Vector Machines*, which were first introduced by Vapnik et al. [Boser, et al. (1992), Vapnik (1995,1998)]. When regarded as a maximization of margin distance that separates different classes, SVM method can be called also „*maximal margin classifier*". The weakness of the term "classifier" is that this method can be also used in regression problems. There are a growing number of publications in this field: Schölkopf (1998), Schölkopf, et al. (1999, 2002), Decoste, et al. (2002), Lin, et al. (2002). Additionally, there are a plenty of Internet resources (http://www.support-vector.net, http://www.kernel-machines.org), where one can find also introduction lectures such as Christianini (2001), Chen, et al. (2001). The aim of this method is to perform some data separation, which simultaneously conforms to tough statistic requirements. This can be achieved by searching for a global minimum of some convex cost function. Since its first publication, the method has been under development and one can find different kinds of it.

The primary task can be defined as follows:

Given a set of linearly separable data (Fig.2)

$$(3) \qquad (\boldsymbol{x}^p, y^p), \quad p = 1,2, \ldots, P, \quad \boldsymbol{x}^p \in \boldsymbol{R}^n, \quad y^p \in \{-1,+1\}.$$

Construct a separating hyper-plane $y(\boldsymbol{x}; \boldsymbol{w}, b)$ so as the separation margin $\delta = 2/\|\boldsymbol{w}\|$ is maximal. Here, $\boldsymbol{w}$, and b are parameters which determine the hyperplane.

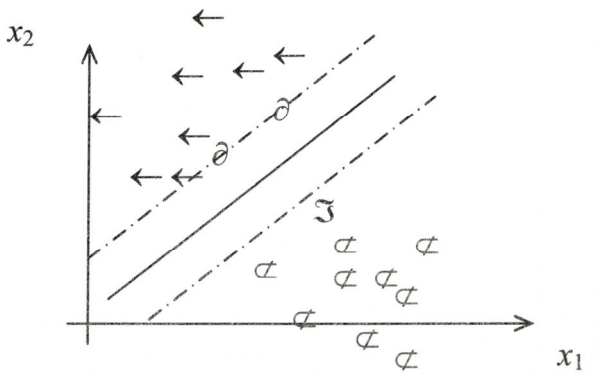

Fig.2. Linearly separable data; two classes

In SVM method, it is suggested that the optimal hyper-plane should be the one which provides maximal (symmetric in relation to separating plane) separation margin of training data. It should be emphasized at this point that the separation is not made using all available data, but only a small part of them (Fig. 2). Intuitively, the separation line should be found basing on the nearest data points.

The way we can maximize the margin δ is the minimization of weight vector norm $\|\boldsymbol{w}\|$ which means the minimization of $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \sum w_i^2$.

Consequently, we should

minimize $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ (4a)

subject to $y^p(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^p + b) \geq 1$ (4b)

Task (4a), (4b) is well known in the optimization theory as square optimization with inequity constraints. Such problems are usually solved by introducing non-negative parameters α^p , the so called Lagrange multipliers and Lagrange function in the form of:

$$(5) \qquad L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{p=1}^{P} \alpha^p [\, y^p(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^p + b) - 1]$$

Providing that $\boldsymbol{\alpha} \geq \boldsymbol{0}$.

The $\tfrac{1}{2}$ coefficient is used for convenience reasons only – it does not break the generality of the given solution. The function must be minimized for $\boldsymbol{w}$ and b and maximized for α^p. It means that we are looking for a saddle point.

It occurs very often that the data are not linearly separable. To ensure the separability of data, some transformation to a higher dimensional space can be introduced. Such a transformation should be subject to Cover theorem [Cover (1965)] about data separability. This theorem says that space can be transformed with high probability to higher multi-dimensional space, in which data are linearly separable. Such a higher-dimensional space is called *feature space.*

It is noteworthy that for such high-dimensional spaces all calculations become extremely difficult. On the other hand, all calculations in SVM method are done using dot products, which means that a specific sort of transformation function can be used, namely *kernel function.*

Definition:
Kernel is a function K such as for all $u,v \in X$:

(6) $$K(u,v) = \langle \phi(u), \phi(v) \rangle$$

where ϕ is the real transform from one space to another.

As stated, the non-linear ϕ transform does not have to be given explicitly. On the other hand, all dot products can be easily calculated in higher-dimensional space using the kernel function, without even knowing exactly what the transform ϕ is.

There are plenty of kernel function types, the most often used being:

Polynomial:

(7) $$K(u,v) = \left(\langle u,v \rangle + 1 \right)^n$$

Gaussian (radial basis):

(8) $$K(u,v) = exp\left(-\frac{1}{2r} \|u - v\|^2 \right)$$

Hyper tangens:

(9) $$K(u,v) = tanh\left(\langle u,v \rangle - \Theta \right)$$

In this paper, two kernel functions have been used: linear (that is polynomial with $n = 1$) and radial with various r values.

The method described above may not work if the data are distorted and not linearly separable even in the feature space. Such data have an apparently good margin, which causes constraints (4b) not to be fulfilled. Also the classes very often overlap each other. When those constraints are not fulfilled, the value of respective Lagrange coefficients is also increased. The way we can accept such distortions of data is to introduce some tolerance for not fulfilling constraints. In other words, we introduce a way to soften the constraints (*soft margin classifiers*) [Cortes, et al. (1995), Cristianini, et al. (2003), Kecman (2001)].

The exceeding value of margins is limited and the limits are set for each example separately. Instead of fulfilling constraints (4b), that is:

$$y^p (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^p + b) \geq 1 \quad (p=1,...,P),$$

one should place the examples within borders:

$$(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^p + b) \geq 1 - \xi^p \quad \text{for} \quad y^p = +1 ;$$
$$(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^p + b) \geq -1 + \xi^p \quad \text{for} \quad y^p = 1.$$

We can formulate two problems depending on the way we consider tolerance $\xi^p \geq 0$ in the quality assessment:

I: minimize $$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{w} + C \sum_p (\xi^p)^2 \qquad (10a)$$

subject to $$y^p (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^p + b) \geq 1 - \xi^p . \qquad (10b)$$

II: minimize $$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{w} + C \sum_p \xi^p \qquad (11a)$$

subject to $$y^p (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}^p + b) \geq 1 - \xi^p , \qquad (11b)$$

$$\xi^p \geq 0 . \qquad (11c)$$

C parameter is used to balance the number of incorrect classifications and the size of the margin – this parameter is set by the author and it remains unchanged during the optimal classifier training. The higher the C value, the bigger the penalty value for errors. As a result, there are fewer wrongly classified points. A smaller C value lets us decrease the margin and tolerate more incorrect classifications.

Diagnosis

Having a sufficient number of training data collected for a particular transformer fault type, one can introduce an expert system able to support the diagnosis of transformers. This would mean creating a classifier which could categorize new data (x, y, z) based on the teaching patterns from a database. Such a classifier would enable one to introduce a more flexible way for aligning classification types with certain transformer types. This seems to be a much better approach than the IEC method.

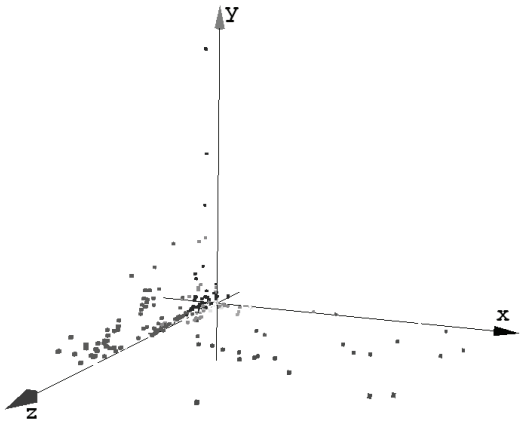

Fig.3. Sample distribution of learning data belonging to nine classes

Because real data collected on power transformers (Fig.3) are distributed in a Cartesian coordinate system, it is difficult to effectively define the boundaries of regions they belong to. Although a part of the data is located in regions that can be quite easily isolated from each other, much of them is grouped and partly mixed near the centre of the coordinate system and one of its axes. Additionally, a part of the decision area contains a small number of data (in practice, some cases occur sporadically and that is why few of them are recorded). What might also cause some difficulty is the disproportion between the numbers of data belonging to each category (Table 3), which is characteristic of this kind of technical problems (in practice, diagnosis is performed on transformers that are heavily damaged). This makes the test even more challenging.

Table 3. Number of measurements classified according to IEC code

Category of technical condition	Number of measurements
No fault	33
Partial disruption of low energy	11
Partial disruption of high energy	2
Full disruption of low energy	10
Full disruption of high energy	20
Overheating below 150 ^{0}C	29
Overheating (150 ^{0}C, 300 ^{0}C⟩	24
Overheating (300 ^{0}C, 700 ^{0}C⟩	79
Overheating by more than 700 ^{0}C	94
Unrecognized fault	64
IEC code not applicable	78
Total:	444

As shown in Table 3, in the majority of cases we observe overheating by more than 300^0C and 700^0C. In a considerable number of cases the diagnosis cannot be made. Partial disruptions of energy are not very frequent. There is also quite a low number of transformers with no faults. This can be explained by the fact that we more often inspect those transformers whose performance is visibly worse, in order to prevent a serious fault or total damage. For verification of the method, 444 measurements from 27 network transformers 250 MVA were available. From this set, the repeating and ambiguous cases must be sorted out. After removing them, there remain 391 examples, 125 out of which being "unrecognized faults".

Application of the classifier based on the IEC reference to the mentioned data yields general correctness rate of about 52%. However, in order to test the ability to create a learning machine, the given set should be split into two sets – one for training and one for testing.

In Table 4 detailed results for classification are presented. In the first column, different types of training datasets are listed and in the remaining columns there are classification results in the form of percentage of correctly labeled objects for different SVM configurations, and IEC for comparison.

The following datasets have been used:

(1) All available points have been used to train the classifier and then it has been tested on the same point set.
(2) Half of the points have been taken for a training set whereas the remaining half - for testing (no repeating data).
(3) 80% point have been used for training, 20% for testing, again no repeating data.

For the configuration, the following abbreviations have been used:

(a) Linear kernel (see 7), C = 10
(b) Linear kernel, C = 50
(c) Radial kernel (see 8), r = 10, C = 50
(d) Radial kernel, r = 50, C = 50
(e) Radial kernel, r = 100, C = 50
(f) IEC

Table 4. Classification results (in %) for different types of datasets

Dataset	(a)	(b)	(c)	(d)	(e)	(f)
(1)	59	60	75	80	80	*57*
(2)	57	57	68	65	63	*58*
(3)	58	61	66	68	70	*53*

It follows from the above that SVM method yields better correctness in comparison to IEC code. Additionally, it emerges that radial basis kernel function allows achieving much better correctness in comparison to linear kernel.

Fig. 4 presents how C coefficient influences the fraction of correctly classified testing vectors in (2) dataset. As x values the C parameter was changed from 10 to 75 with step 10, as y there are fractions of correctly classified vectors ($y \in [0,1]$).

In fact, changing the penalty factor value does not influence much the classification correctness. The threshold value is about 30, after which there are virtually no changes in correctness.

Because in SVM classifiers with radial kernel have definitely yielded better results, one should examine how exactly that kernel works. Fig. 5 presents the dependencies between radius value in the radial kernel function and the yielded correctness in (2) dataset. On the x axis there is r value from formula (8) changed from 5 to 75 with step 5, the y axis contains again fraction of correctly classified vectors. The lower series depicts correctness achieved on a testing set.

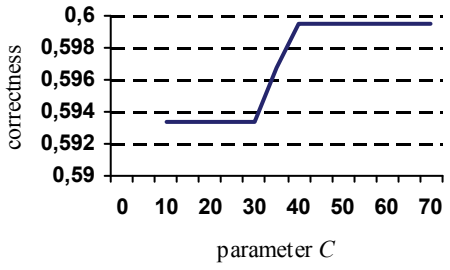

Fig 4. The influence of *C* on the classification's correctness

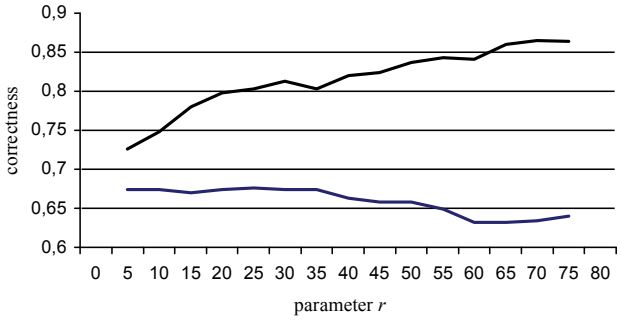

Fig 5. The influence of *r* on the classification's correctness

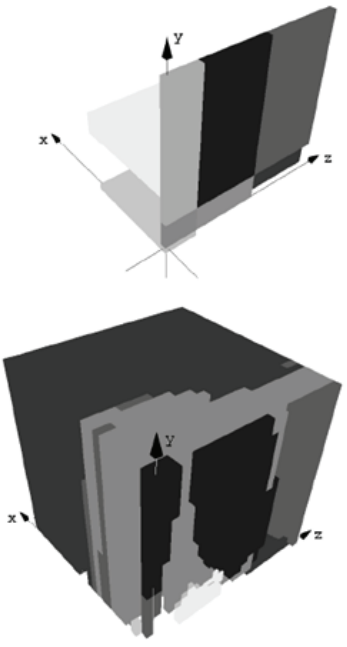

Fig 6. IEC (top) versus SVM (bottom) results

Obviously, the higher *r* value, the better is the correctness. On the other hand, higher *r* causes the classifier to be "overtrained", which is typical for a pattern learning process. It means that *r* value must be a compromise between good correctness on training set and the ability to generalize knowledge, which in turn is depicted by correctness on a testing set.

Fig. 6 depicts the regions created by SVM as compared to the IEC ones. One can easily notice the difference: while IEC leaves most of the feature space unrecognized, SVM

classifier introduces classification in the whole space by extrapolating the rules learnt from the training patterns.

Eventually, the SVM method enables one to create a classifier which is capable of introducing feature space separation that is about 15 – 20 percent more accurate than the standard IEC approach

Summary

The maximal margin classifier enables the creation of an expert system which diagnoses technical condition of an oil transformer basing on results of chromatography of gases dissolved in transformer oil (Dissolved Gas Analysis – DGA). The system is designed to diagnose small and medium size units for which DGA is an essential source of information on their technical condition. The diagnosis is based on the separation of learning data, which enables the creation of regions reflecting the same state of the transformer. On that basis, new results of DGA can be examined more accurately than in the case of the IEC method.

Therefore, a new method for transformer diagnosis support has been developed. The method, based on real measure data collected from a particular category of transformers, reflects in a way the process of learning from experience. Of course, the implementation of the presented system in industry will demand much further research and testing on numerous learning sets, particularly for the categories that are rarely recorded in practice. The method presented can be also employed in other tasks of grouping and classification of hardly separable data.

REFERENCES

[1] IEC (1979): International Electrotechnical Commission, Interpretation of the Analysis of Gases in Transformers and other Oil-Filled Electrical Equipment in Service. Geneva, 1979.

[2] B.E.Boser, I.M.Guyon, V.N.Vapnik (1992): A training algorithm for optimal margin classifier. In: D.Haussler (Ed.): 5th Annual ACM Workshop on COLT. Pittsburg, PA, ACM Press, 144-152.

[3] V.N.Vapnik (1995): The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, 1995.

[4] V.N.Vapnik (1998): Statistical Learning Theory. Wiley, New York.

[5] N.Cristianini (2001): ICML'01 Tutorial. http://www.kernel-machines.org .

[6] N.Cristianini, J.Shawe-Taylor (2003): Support Vectors and Kernel Methods. In: M.Berthold, D.J.Hand (Eds.): Intelligent Data Analysis; An Introduction. Springer-Verlag, Berlin, Heidelberg.

[7] P.-H.Chen, C.-J.Lin, B.Schölkopf (2001): A tutorial on □-Support Vector Machines. http://www.kernel-machines.org .

[8] C.Cortes, V.N.Vapnik (1995): Support Vector Networks. Machine Learning, 20, 273-297.

[9] T.M.Cover (1965): Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. IEEE Trans. on Electronic Computers, 14, 326-334.

[10] D.Decoste, B.Schölkopf (2002): Training Invariant Support Machines. Machine Learning, 46, 161-190.

[11] V.Kecman (2001): Learning and Soft Computing. The MIT Press, Cambridge, Massachusetts.

[12] B. Schölkopf, C.J.C.Burges, A.J.Smola (Eds.): Advances in Kernel Methods: Support Vector Learning. The MIT Press, Cambridge, MA; 255-268.

[13] Y.Lin, Y.Lee, G.Wahba (2002): Support vector machines for classification in nonstandard situations. Machine Learning, 46, nos.1-3; 191-202.

[14] B. Schölkopf (1997): Support vector learning. R.Oldenbourg Verlag, München. Doktorarbeit, TU Berlin; http://www.kernel-machines.org

[15] B.Schölkopf, A.Smola (2002): Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, 2002.

[16] M.Ulewicz (2003): Support Vector Machines with Distortions for Handwritten Digits Recognition. In: M.Kurzyński, E.Puchała, M.Woźniak (Eds.), KOSYR'2003 – Computer Recognition Systems, Wrocław, 109-114.

***Authors**: prof. dr hab. inż. Piotr S. Szczepaniak, Technical University of Łódź, Institute of Information Technology, ul. Wólczańska 215, 90-924 Łódź, Poland, E-mail: piotr@ics.p.lodz.pl; mgr inż. Marcin Kłosiński, Technical University of Łódź, Institute of Information Technology, ul. Wólczańska 215, 90-924 Łódź, Poland, E-mail: klosinskim@tt.com.pl;*