The regularization method in the classification task according to given examples

Włodzimierz KWIATKOWSKI

Institute of Teleinformatics and Cybersecurity, Faculty of Cybernetics, MUT,
ul. gen. Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland
wlodzimierz.kwiatkowski@wat.edu.pl

ABSTRACT: The article considers the problem of classification based on the given examples of classes. As a feature vector, a complete characteristic of object is assumed. The peculiarity of the problem being solved is that the number of examples of the class may be less than the dimension of the feature vector, and also most of the coordinates of the feature vector can be correlated. As a consequence, the feature covariance matrix calculated for the cluster of examples may be singular or ill-conditioned. This disenable a direct use of metrics based on this covariance matrix. The article presents a regularization method involving the additional use of statistical properties of the environment.

KEYWORDS: regularization, classification, pattern recognition, exploratory data analysis.

1. Introduction

The methods presented in this article apply to the tasks of classifying objects based on their features in the form of real number vectors. The solution proposed can be used especially when a feature vector is defined as a complete characteristic of objects, rather than previously defined attributes. This usually happens when the classification is based on automatically collected data (for example – measurement results), without selection from the point of view of discriminatory properties. This requires an analysis of vectors of large dimensions and large variety. In this case, mining methods, commonly referred to as exploratory data analysis, show promise for the future.

Methods for determining classification rules examined in this article are based on comparing the distance of clusters composed of the given examples of classes from the feature vector of the analyzed object. The size of an example

cluster is usually small – compared to the dimension of the feature space. This poses a significant problem when the classification is based on the metrics defined separately for individual example clusters.

The classification task on the basis of distances defined separately for each class is presented in paper [7]. This approach refers directly to quadratic discriminant analysis (QDA). Where the feature covariance matrix of example cluster is singular, this approach leads to the concept of using the generalised Mahalanobis distance [12]. This concept is based on the Moore-Penrose pseudo-inverse of covariance matrix. However, solutions based on such approach may turn out to be completely wrong.

The method proposed in this article involves formulating the derivative classification task. This task is formulated for the case when the pattern feature covariance matrices are singular or ill-conditioned (there is a large range between their eigenvalues and their determinants are close to zero). The derivative task is constructed to eliminate the reason for not obtaining an unambiguous solution. This is achieved by supplementing the original task with additional information. In the problem under consideration, this is realized by supplementing the distance function – based on the statistical properties of the example cluster – with a regularization term based on the statistical properties of the environment. The presented approach is interpreted as a method for regularizing the original classification task.

The problem of regularizing classification has been studied in various applications and from various points of view [3], [5], [11] and [14]. The following issues are related to the approach discussed in this article.

Analysis of data from many sources is one of the important problems associated with classification. In most cases, such data cannot be modelled by a common, multidimensional statistical model. Methods based on various models and setting the rules for obtaining a compromise solution are used in this case. Here, we will cite paper [1] as an example, which presents consensus theory-based methods. The use of regularization is one of the conditions for obtaining compromise solutions.

The problem of cooperation with the decision-maker (user) to obtain compromise solutions is a separate topic. The method discussed in this article employs the knowledge (experience) of the decision-maker given by indicating examples of patterns [8], [9]. Similar issues occur in the problems of semi--supervised learning algorithms [6], [15], [16], which combine labelled (marked) and unlabelled data. These algorithms are gaining significant interest and are successfully implemented in practical applications for data mining [13], [14]. In these algorithms, the problem of regularization is also significant, and its solution usually involves the idea of penalization [3].

The issue of classification based on the assessment of distance between clusters in the feature space is presented in papers [2], [7]. An example of using such functions is presented in [4].

2. Formulation of the classification problem based on given examples

The given set of objects is numbered 1 to N. A feature vector expressed in real numbers is known for each object. We use the following designation for object number k:

$$\mathbf{a}_k = \left[a_{1,k}, \quad a_{2,k}, \quad \cdots \quad a_{L,k}\right]^T, \qquad \mathbf{a}_k \in R^L \tag{1}$$

Each coordinate $a_{l,k}$ is a real number and parameter L determines the number of vector coordinates. These vectors form a set:

$$\mathrm{A} = \left\{\mathbf{a}_1, \quad \mathbf{a}_2, \quad \ldots \quad \mathbf{a}_N\right\}, \qquad \mathbf{a}_k \in R^L \tag{2}$$

The feature vectors are compiled as the following matrix:

$$\mathbf{A} = \left[\mathbf{a}_1, \quad \mathbf{a}_2, \quad \ldots \quad \mathbf{a}_N\right], \qquad \mathbf{a}_k \in R^L \tag{3}$$

The feature vectors covariance matrix is determined as follows:

$$\mathbf{R} = \tfrac{1}{N-1}\sum_{k=1}^{N}\left(\mathbf{a}_k - \overline{\mathbf{a}}\right)\left(\mathbf{a}_k - \overline{\mathbf{a}}\right)^T \tag{4}$$

where:

$$\overline{\mathbf{a}} = \tfrac{1}{N}\sum_{k=1}^{N}\mathbf{a}_k \tag{5}$$

It is then assumed that

$$\det(\mathbf{R}) \neq 0 \tag{6}$$

The distance between vectors $\mathbf{x}$, $\mathbf{y}$ of feature space R^L is determined in a way that takes into account the dispersion of coordinates and their mutual correlation. This requirement is met by the Mahalanobis distance [10]. It is set by the formula:

$$\mathrm{d}_{\mathrm{e}}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\,\mathbf{R}^{-1}(\mathbf{x}-\mathbf{y})}\,, \qquad \mathbf{x},\mathbf{y} \in R^L \tag{7}$$

Examples constituting the class pattern with index $h \in \{1,2,\ldots,H\}$ (where: H – number of classes) are indicated by providing the relevant set of

indexes W_h. Class pattern with index h is therefore represented by the following set of points (cluster) in the feature space:

$$C(W_h) = \{\mathbf{w}_k \in \mathrm{A} : k \in W_h\} \tag{8}$$

The number of elements of such pattern W_h is marked as $N_h = \|C(W_h)\|$.

Inference about the similarity of feature $\mathbf{x}$ to pattern W_h is based on the distance of point $\mathbf{x}$ from cluster $C(W_h)$. For example, the choice of centroid method to determine the distance between clusters results in:

$$\mathrm{D_e}(\mathbf{x}, C(W_h)) = \mathrm{d_e}(\mathbf{x}, \overline{\mathbf{w}}_h) = \sqrt{(\mathbf{x} - \overline{\mathbf{w}}_h)^T \mathbf{R}^{-1} (\mathbf{x} - \overline{\mathbf{w}}_h)} \tag{9}$$

where:

$$\overline{\mathbf{w}}_h = \tfrac{1}{N_h} \sum_{j \in W_h} \mathbf{w}_j \tag{10}$$

The classification based on the metric (9) is called environmental.

Due to the method of determining the covariance matrix $\mathbf{R}$, the use of environmental classification is justified when the features of all patterns are uniform in the following sense: the relevant clusters differ only in expected values (and the corresponding covariance matrices are the same). If the pattern covariance matrices differ, it is recommended to diversify the way the distances are measured according to the covariance matrices of respective patterns [7].

Covariance matrix based on examples of pattern W_h are marked as follows:

$$\mathbf{R}_h = \tfrac{1}{N_h - 1} \sum_{j \in W_h} \left(\mathbf{w}_j - \overline{\mathbf{w}}_h\right)\left(\mathbf{w}_j - \overline{\mathbf{w}}_h\right)^T \tag{11}$$

Distance between vectors $\mathbf{x}$, $\mathbf{y}$ of feature space R^L is matched to the pattern W_h, if it is expressed by the formula [7]:

$$\mathrm{d}_h(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{R}_h^{-1} (\mathbf{x} - \mathbf{y})}, \qquad \mathbf{x}, \mathbf{y} \in R^L \tag{12}$$

We similarly refer to the distance between feature $\mathbf{x}$ and cluster $C(W_h)$. For example, the distance is specified by the following formula for the centroid method:

$$\mathrm{D}_h(\mathbf{x}, C(W_h)) = \mathrm{d}_h(\mathbf{x}, \overline{\mathbf{w}}_h) = \sqrt{(\mathbf{x} - \overline{\mathbf{w}}_h)^T \mathbf{R}_h^{-1} (\mathbf{x} - \overline{\mathbf{w}}_h)} \tag{13}$$

The classification based on metrics (12) suitably matched to individual patterns is referred to as the classification matched to patterns. The usefulness of such a classification, which means differentiating the method of distance

calculation according to the pattern covariance matrix, is illustrated by the example in Figure 1. The example applies to the division of space R^2 into two classes based on given patterns: W_1 and W_2. Points $C(W_1)$ are shown in the figure as circles, points $C(W_2)$ – as squares. Feature space points closer to points $C(W_1)$ they are marked in a darker colour. In the example, clusters $C(W_1)$ and $C(W_2)$ are not linearly separable and the environmental classification gave poor results. The results of classification matched to the patterns are as expected.

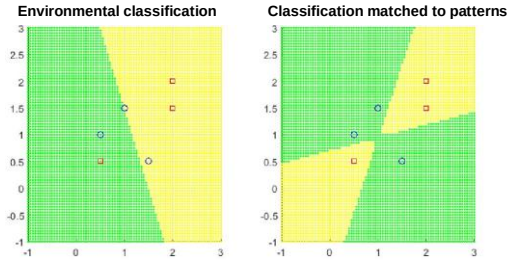

Fig. 1. Example of classification when the pattern covariance matrices are non-singular and the formula (13) is applied. On the left: the distance is determined using a metric based on the covariance matrix calculated for all patterns together, as per the formula (4). On the right: the distance is determined using metrics based on covariance matrices calculated separately for the features of each pattern, as per the formula (11)

3. The method for regularizing the classification task

The problem solved in this article applies to regularizing the task of classification matched to patterns. Regularization is needed when pattern covariance matrices $\mathbf{R}_h$ are singular or ill-conditioned. In the discussed problem, ill-conditioning is understood as a very wide range between the eigenvalues of matrix $\mathbf{R}_h$, causing the matrix determinant to be close to zero.

A routine procedure in the case presented is the application of the generalised Mahalanobis distance, defined as follows [12]:

$$\mathrm{D}_h(\mathbf{x}, C(W_h)) = \mathrm{d}_h(\mathbf{x}, \overline{\mathbf{w}}_h) = \sqrt{(\mathbf{x} - \overline{\mathbf{w}}_h)^T \mathbf{R}_h^+ (\mathbf{x} - \overline{\mathbf{w}}_h)} \tag{14}$$

where: $\mathbf{R}_h^+$ - Moore-Penrose pseudo-inverse of the covariance matrix $\mathbf{R}_h$.
However, in classification tasks based on patterns that are not separable linearly, the application of generalised Mahalanobis distance may lead to false solutions. This is illustrated by the example in Figure 2. As in the example above, space R^2 is divided into two classes based on given patterns: W_1 and W_2. Points

$C(W_1)$ are shown in the figure as circles, points $C(W_2)$ – as squares. Compared to the previous example, both clusters W_1 and W_2 are less numerous: each class is indicated by only two examples. The points of the feature space closer to points $C(W_1)$ are marked in a darker colour. In the example, clusters $C(W_1)$ and $C(W_2)$ are not linearly separable and both classification methods have bad (unexpected) results, and the result of the method using matched metrics is quite the opposite of what was expected.

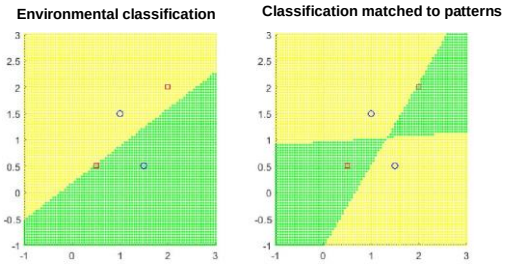

Fig. 2. Example of classification when the pattern covariance matrices are singular and the formula (14) is applied. On the left: the distance is determined using the metrics based on the covariance matrix calculated for all patterns together, as per the formula (4). On the right: the distance is determined using metrics based on covariance matrices calculated separately for each pattern, as per the formula (11)

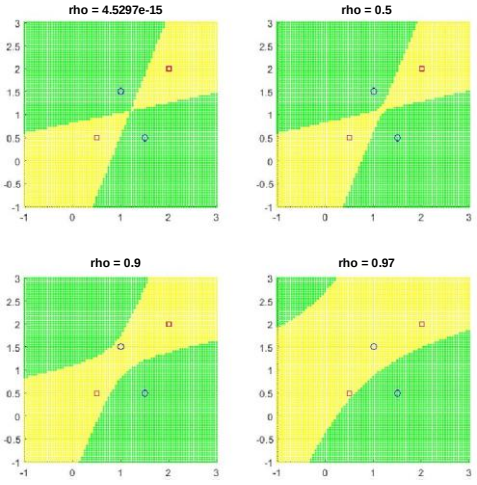

Fig. 3. Illustration of the results of matched classification using regularization

We base the proposed method of regularization on the introduction of a distance function whose values are defined as follows:

$$\mathrm{d}_h^r(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T[(1-\rho)\mathbf{R}_h + \rho\mathbf{R}]^{-1}(\mathbf{x}-\mathbf{y})}\,, \quad \mathbf{x},\mathbf{y} \in R^L \tag{15}$$

where: $\rho \in [0,1]$ – regularization parameter.

Regularization consists in replacing the covariance matrix $\mathbf{R}_h$ with a convex combination of matrix $\mathbf{R}_h$ and matrix $\mathbf{R}$. Value $\rho = 0$ means no regularization and matching classification, while value $\rho = 1$ means transition to the environmental classification.

The results are illustrated for the data as in Figure 2. Figure 3 shows the results of matched classification using regularization. Correct classification results have already been observed for the value approx. 10^{-15} of the regularization parameter. The maximum value of this parameter was approx. 0.5 (a further increase in the parameter causes a smooth transition to the results of the environmental classification).

Figure 4 presents similar results for the task of dividing the feature space into three classes. The comparison included the results of the classification based on Euclidean metric (in the figure marked as Euclidean classification), the classification based on the Mahalanobis metric (in the figure marked as environmental classification), the pattern-matched classification, based on the generalised Mahalanobis metric (marked as matched classification) and pattern-matched classification using regularization with parameter $\rho = 0{,}01$ (marked as regularized classification). We can see that only the results of the last classification gave satisfactory results.

To illustrate the impact of the regularization parameter on the quality of classification, we present the result of a computational experiment consisting in dividing a set of objects into two classes. Features of N first class objects and the same number of second class objects have been randomised in the experiment. Of these, N_1 examples of first class objects and N_2 examples for second-class objects have been indicated at random. For both classes of objects, the features are points on the plane, randomised according to properly selected normal distributions. The following are assumed in the example in Figure 5: $N = 20$, $N_1 = 4$, $N_2 = 2$. It can be seen that the subspace generated by two object indications is a straight line and the matched classification task is ill-conditioned. Regularization of this task at parameter $\rho = 0{,}01$ has allowed us to obtain correct classification results.

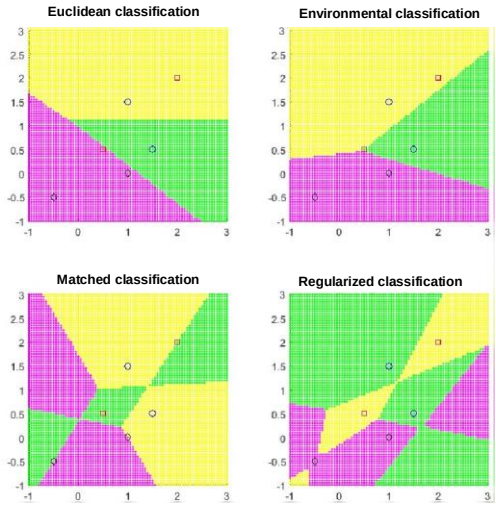

Fig. 4. Comparison of regularized classification results with classification results obtained through other methods

Collective results of the computational experiment under discussion are presented in Figure 6. The abscissa axis indicates the rate of misclassification to class 2, and the ordinate axis – the rate of misclassification to class 1. These rates were determined based on 1000 tests. The green colour indicates the classification results using the matched regularized method for various values of regularization parameter $\rho \in (0,1)$. The end point for the value $\rho = 1$ (marked in blue) corresponds to the quality of the environmental classification. The end point for the value $\rho = 0$ (marked in red) corresponds to the quality of the matched classification without regularization. There is a noticeable lack of continuity of features when transiting from zero value of the regularization parameter ($\rho = 0$) to a positive value. The leap observed in the experiment occurred at the value $\rho \approx 10^{-14}$. This is the minimum value of the regularization parameter for the computing environment being used. The obtained quality curves for regularized classification tend to form a curve illustrating the situation when regularization is not necessary ($N_1 = 4$, $N_2 = 4$). However, also in this case, it is possible to slightly improve the classification quality through regularization. In the experiment discussed, acceptable classification results have been obtained for the regularization parameter value of 0.01 to 0.1.

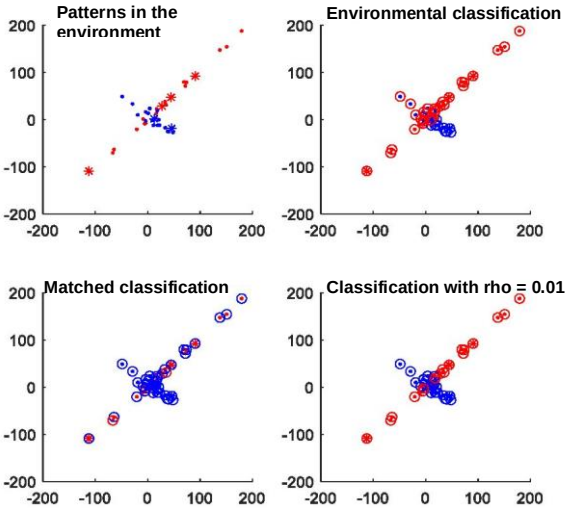

Fig. 5. Example of a computational experiment. Class 1 objects and class 2 objects are marked with red and blue points, respectively. The examples are marked with stars of the relevant colour. Classification results are marked with circles of the relevant colour

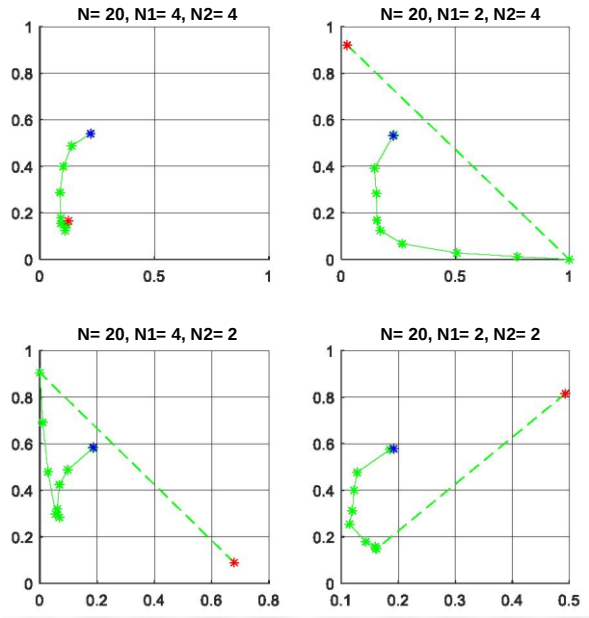

Fig. 6. Illustration of the impact of regularization parameter on the classification quality. The abscissa axis shows the rate of misclassification to class 2, and the ordinate axis - the rate of misclassification to class 1

4. Conclusions

1) The proposed method for regularization can be applied wherever the features of classified objects can be presented as vectors of real numbers. If the covariance matrix of all examined objects is singular (or ill-conditioned), pre-processing should be carried out to select the features that will ensure the non-singularity of their covariance matrix.
2) The interpretation of the derivative task of classification is clear. The proposed approach consists in supplementing the pattern data with data on statistical properties of the environment.
3) The calculation algorithm is attractive because of its simplicity. It allows the use of more complex methods for determining the distance between clusters than method applied in the examples shown in the article. It is also possible to obtain classifications based on different rules for assessing cluster similarity.
4) Classification resulting in trivial, multiple or too extensive classes usually means inconsistencies in the indicated class patterns.

Literature

[1] BENEDIKTSSON J.A., BENEDIKTSSON, K., *Hybrid consensus theoretic classification with pruning and regularization*. IEEE 1999 International Geoscience and Remote Sensing Symposium, 1999, Volume 5, pp. 2486-2488.

[2] FUKUNAGA, K., *Feature Extraction Algorithm Using Distance Transformation.* IEEE Transactions on Computers C-21(1), February 1972, pp. 56-63.

[3] HSUN-HSIEN CHANG, MOURA JOSE M. F., *Classification by Cheeger Constant Regularization.* 2007 IEEE International Conference on Image Processing, 2007, Volume 2, pp. II-209-II-212.

[4] JIANGTAO PENG, LEFEI ZHANG, LUOQING LI, *Regularized set-to-set distance metric learning for hyperspectral image classification*. Pattern Recognition Letters, Volume 83, Part 2, 1 November 2016, pp. 143-151.

[5] JIM JING-YAN WANG, YI WANG, SHIGUANG ZHAO, XIN GAO, *Maximum mutual information regularized classification.* Engineering Applications of Artificial Intelligence, Volume 37, January 2015, pp. 1-8.

[6] JUN WANG, GUANGJUN YAO, GUOXIAN YU, *Semi-supervised classification by discriminative regularization*. Applied Soft Computing, Volume 58, September 2017, pp. 245-255.

[7] KWIATKOWSKI W., *Metody automatycznego rozpoznawania wzorców*. BEL Studio, Warszawa, 2010.

[8] KWIATKOWSKI W., *Wykrywanie anomalii bazujące na wskazanych przykładach.* Przegląd Teleinformatyczny, nr 1-2, 2018, s. 3-21.

[9] KWIATKOWSKI W., *Recommendations as a result of decision evaluations based on reference examples.* Teleinformatics Review, No. 1-2, 2019, pp. 3-23.

[10] MAHALANOBIS P.C., *On the generalized distance in statistics.* Proceedings of National Institute of Sciences (India), Vol. 2, No. 1, 1936, pp. 49-55.

[11] MAJUMDAR A., WARD R. K., *Classification via group sparsity promoting regularization*. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 861-864.

[12] WARMUS M., *Uogólnienie odległości Mahalanobisa*. Listy Biometryczne, Nr 30-33, 1971, s. 3-7.

[13] TAO ZHANG, CHEN GONG, WENJING JIA, XIAONING SONG, JUN SUN, XIAOJUN WU, *Supervised Image Classification with Self-Paced Regularization*. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 411-414.

[14] YANG LI, DAPENG TAO, WEIFENG LIU, YANJIANG WANG, *Co-regularization for classification*. 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2014, pp. 218-222.

[15] YATING SHEN, YUNYUN WANG, ZHIGUO MAARMUS, *Label-expanded manifold regularization for semi-supervised classification*. 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2017, pp. 1-4.

[16] ZHILEI CHAI, WEI SONG, HUILING WANG, FEI LIU, *A semi-supervised auto-encoder using label and sparse regularizations for classification*. Applied Soft Computing, Volume 77, April 2019, pp. 205-217.

Metoda regularyzacji w zadaniu klasyfikacji według zadanych przykładów

STRESZCZENIE: W artykule rozpatrywany jest problem klasyfikacji na podstawie wskazanych przykładów klas. Jako wektor cech przyjmuje się kompletną charakterystykę obiektów. Osobliwość rozwiązywanego zadania wynika z tego, że liczba przykładów klasy może być mniejsza od wymiaru wektora cech, a także wektor cech może zawierać współrzędne skorelowane. W konsekwencji macierz kowariancji cech obliczana dla klastra przykładów może być osobliwa albo źle uwarunkowana. Uniemożliwia to bezpośrednie stosowanie metryk bazujących na tej macierzy kowariancji. W artykule została przedstawiona metoda regularyzacji polegająca na dodatkowym wykorzystaniu statystycznych właściwości środowiska.

SŁOWA KLUCZOWE: regularyzacja, klasyfikacja, rozpoznawanie wzorców, eksploracja danych

Received by the editorial staff on: 29.04.2019