DOI: 10.5604/01.3001.0010.4853

TESTING FOR REVEALING OF DATA STRUCTURE BASED ON THE HYBRID APPROACH

Volodymyr Mosorov, Taras Panskyi, Sebastian Biedron
Lodz University of Technology, Institute of Applied Computer Science

Abstract. *In this paper testing for revealing data structure based on a hybrid approach has been presented. The hybrid approach used during the testing suggests defining a pre-clustering hypothesis, defining a pre-clustering statistic and assuming the homogeneity of the data under pre-defined hypothesis, applying the same clustering procedure for a data set of interest, and comparing results obtained under the pre-clustering statistic with the results from the data set of interest. The pros and cons of the hybrid approach have been also considered.*

Keywords:– pre-clustering hypothesis, data group structure testing, group structure revealing

TESTOWANIE WYSTĘPOWANIA STRUKTURY DANYCH W OPARCIU O PODEJŚCIE HYBRYDOWE

Streszczenie. *W pracy tej przedstawiono testowanie występowania struktury danych w oparciu o podejście hybrydowe. Podejście to, podczas testowania wymaga zdefiniowania hipotezy wstępnego klastrowania; założenia homogeniczności danych na podstawie zdefiniowanej „statystyki"; zastosowania tej samej procedury klastrowania dla interesującego zbioru danych oraz porównania wyników uzyskanych na podstawie statystyki z wynikami uzyskanymi z interesującego nas zbioru danych. Zalety i wady podejścia hybrydowego zostały również rozważone*.

Słowa kluczowe: hipoteza wstępnego klastrowania, testowanie struktury danych, występowania struktury

Introduction

Cluster analysis serves as a tool for finding groups of objects in data sets. There are three notoriously hard problems in cluster analysis: the estimation of a number of clusters, checking whether the data set to be clustered is actually homogeneous or not and the validation of the clustering results.

When clustering operation is applied to a set of data, objects in it are classified whether or not the data exhibit a true or natural grouping structure. If the unknown structure of the data is to be discovered, artificial clustering is not acceptable. Moreover, if the data do not possess any group structure and clustering is applied without the testing for absence/presence of data structure, the output of a clustering can provide a misleading group indication. Therefore, the logical starting point for a cluster analysis must be a test for revealing the group structure.

In most applications, technical systems performing a clustering operation do not analyze its necessity. Thus, it is typical in automatic visual inspection systems [1], automatic event detection from video sequences [2], speaker clustering aided by visual dialogue analysis [3]. The performance of the clustering operation is erroneous when the data represent one cluster, while the technical system is "forced" for the automated clustering. Therefore, testing for revealing the data structure is the main objective of this paper.

The hybrid approach used during the testing process at first suggests defining a pre-clustering *hypothesis*. The second step is to define a pre-clustering *statistic* and to assume the homogeneity of the data under pre-defined *hypothesis*. Then the same clustering procedure should be applied to a data set of interest and results obtained under the *statistic* are to be compared with the results from the data set of interest.

Since testing for revealing the data structure is the main objective of this article, the problem of choosing the attribute space and the problem of discovering the optimal number of clusters will not be considered.

1. Related works

The following validation tests have been proposed to reveal the group structure. Tests of the Poisson model have been based on several assumptions, namely, the number of pairwise distances less than a specified threshold [4] or the largest nearest neighbor distance within a set of objects [5]. Some tests are based on the predefined thresholds or indicators searching for "gaps" in a data under the Poisson and unimodal models [6]. Alternative approaches specifying the presence of a structure in the data have been presented in [7] and [8]. These tests are based on the comparison of the calculated coefficients obtained from the original data and those from artificial pseudo-generated homogeneous data in conjunction with the validity of clustering indices.

Tests for revealing the group structure in the data are not usually employed in practical clustering applications [9]. As it was noticed in [10], this situation is caused by the users' strong presumption that their data do contain group structure. However, most tests that use the sequential calculation of a coefficient assuming homogeneity of data and thereby forming a set of statistics are based on the users' specified thresholds while others are tailored mainly to the normal density situations with globular clusters of the same variance. Moreover, some of these tests use a number of assumptions about the form of the cluster, objects distribution or data type. However, the biggest problem is ignoring the tests and increasing attention that is paid in recent years to providing ways of testing the number of clusters that are more formal and obvious than tests described above. Nevertheless, the subject has some practical interest and, therefore, the test for revealing the data structure remains an integral part of clustering procedure which cannot be omitted.

2. The general setup

The set of n d-dimensional data objects $X = \{x_1, \ldots, x_n\}$ in Euclidean space P^d is given, where $X \in P^d$ and each object belongs to the set Ξ, where Ξ is a non-empty, finite set called the *data*. Then there is a clustering method C, so that $C(X,k) = \{C_1, \ldots, C_k\}$ with $k \in K$, and, for $i = 1, \ldots, k : C_i \subseteq X$. In our case, C is a partitioning method assuming that $C_i \bigcap C_j = \varnothing$ for any $i \neq j$. Furthermore, a set of p values in S (pre-clustering statistic) is given, so that $S(X, C(x,k)) \in P^d$ under pre-clustering hypothesis T_1 serves as an additional parameter for validation purpose (see Section 3).

For a fixed number of clusters (default $k = 2$) and fixed objects in the data set X, a test for revealing the data structure is defined by estimating the $p_{obs}{}^*$-value.

$$p_{obs}{}^* = \begin{cases} negative, & D - (d_1 + d_2) < 0 \\ positive, & otherwise \end{cases}, \tag{1}$$

where D is average pairwise distance within a data set X and it could be estimated as follows:

$$D = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2}{n(n-1)}} \qquad i, j = 1..n \qquad i \neq j. \tag{2}$$

d_1 and d_2 are average pairwise distances within clusters C_1 and C_2 respectively and are computed as follows:

$$d_p = \sqrt{\frac{\sum_{i=1}^{n_p}\sum_{j=1}^{n_p}(x_i - x_j)^2}{n_p(n_p-1)}} \qquad i, j = 1..n; \quad i \neq j; \quad p = 1, 2 \tag{3}$$

In practice, for convenience, the categorical (qualitative) value of $p_{obs}{}^*$ could be changed to the numerical (quantitative) value p_{obs} and written in the form of real numbers.

If $p_{obs}{}^* = negative$ or $p_{obs} < 0$, and $p_{obs} \in S$, this shows the homogeneity of data *X*, which therefore suggests that clustering needn't be performed. Otherwise, the data have true, natural group structure and the use of the clustering operation is justified.

3. Testing for manifestation of data structure

Pre-clustering hypothesis

For running the test for revealing the data structure, pre-clustering hypotheses need to be determined. The basis of a test is a statement of "no structure" or data randomness that is called a pre-clustering hypothesis. This requires researchers' judgment, because it depends on what constitutes "homogeneity of data" in the given application.

Let $T_1 = \{y_1, \ldots, y_n\} \in P^d$ be n d-dimensional observations which represent *n* data objects under investigation and which are to be analyzed assuming the homogeneity of group structure. Probabilistic models should be used and therefore $y_1, \ldots, y_n$ are considered as variables of *n* independent *d*-dimensional random vectors $Y_1, \ldots, Y_n$ of unimodal distribution *f(y)*, where mode of a distribution attains its maximum. In Figure 1 a homogeneous sample affirming the pre-clustering hypothesis has been depicted.

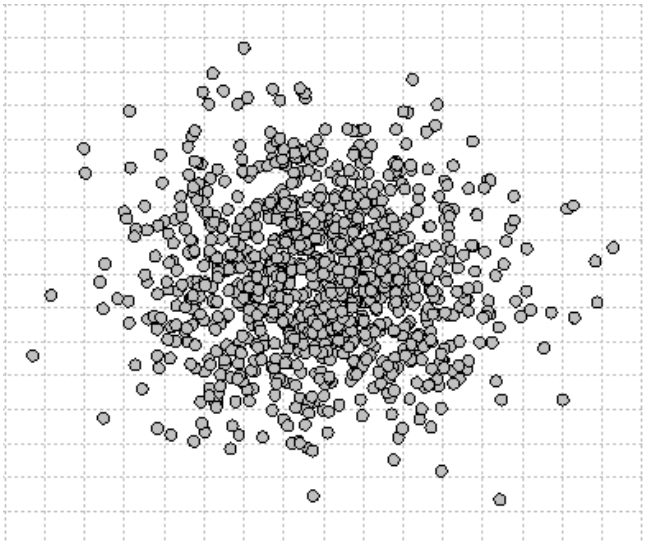

Fig. 1. Data sample under the pre-clustering hypothesis

The underlying assumption of the pre-clustering hypothesis is that there is only one general cluster in the data set. Furthermore, this hypothesis allows for the possibility that data objects are closer to one another when they are located nearer to the center of the cluster, then they are on the boundary, being thus in practice similar to normal distribution as it is shown in Figure 1.

The pre-clustering hypothesis has to be defined in such a way that its parameters can be easily estimated by the different types of data (numerical, categorical, binary) using different partitioning clustering algorithms (*k*-medoid, *k*-kernel, *k*-median, etc.).

Pre-clustering statistic

If pre-clustering hypothesis T_1 is already defined, next step is to determine a model, which assumes the data homogeneity. In statistical modeling, it is common practice that models are created after a study of data, and then, for the adequacy of these models, they are assessed on different similar data sets which are believed to have the same properties as the original one. With this in mind, pre-clustering statistic *S* under pre-clustering hypothesis is determined as follows:

1) Calculate the *p*-value based on the pre-clustering hypothesis T_1.
2) Calculate the same *p*-value for a large number (say, N_{rep}) of data sets simulated independently under the pre-clustering hypothesis of no clustering and let these values be $S = \{T_1, T_2, \ldots, T_{N_{rep}}\}$. A number of simulated data sets N_{rep} is often purely heuristic. Data sets have been simulated according to normal unimodal distribution with a change in the standard deviation ranging $\{\sigma_{\min}, \sigma_{\max}, d\sigma\}$ where $\sigma_{\min}$ is an arbitrarily positive value, however small, but not zero; $\sigma_{\max}$ is an arbitrarily positive value that in theory tends to infinity, however in this article it is limited to $\sigma_{\max} = 2$, $d\sigma$ is a value that represents a size of a step. A graphical representation of pre-clustering statistic is shown in Figure 2.

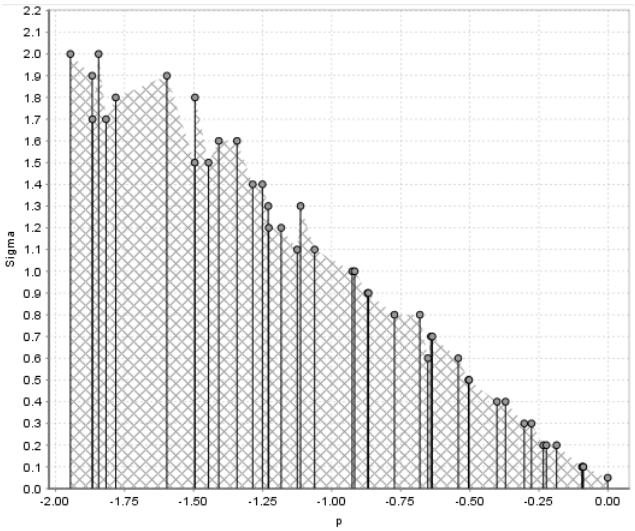

Fig. 2. Pre-clustering statistic

From Figure 2 it can be concluded that standard deviation increases linearly in proportion to the reduction of *p*-value. It should also be noted that if the pre-clustering statistic is already determined, it can be used as a *template* for testing various data sets.

The pre-clustering statistic is determined by repeated estimation of *p*-values for a data distributed according to Gauss's law that form one globular cluster. The Gaussian cluster has been repeatedly generated with different values of standard deviation σ to form a statistic and thus, *p*-value's parameters, namely d_1, d_2, and *D* have been estimated under *k*-means clustering (always $k = 2$) separately for each repetition.

Processing steps

- Run the *k*-means clustering algorithm for a data set of interest, with $k = 2$. Estimate d_1, d_2 and *D*, within a cluster C_1, C_2 and *X* respectively.
- Define the pre-clustering statistic *S* under the pre-clustering hypothesis T_1 or use the already defined statistic as a predesigned template.
- Compare p_{obs} – value obtained from the set of interest with a statistic *S* which indicates pre-clustering hypothesis and thus reveals data structure.

The calculation of each individual *p*-value, while forming the *S* pre-clustering statistic, is made according to formula (1). The difference between p and p_{obs} values is that the first is calculated to form the statistic, whereas the second is calculated for the set of interest. The proposed approach for testing for revealing the data group structure is shown in Figure 3.

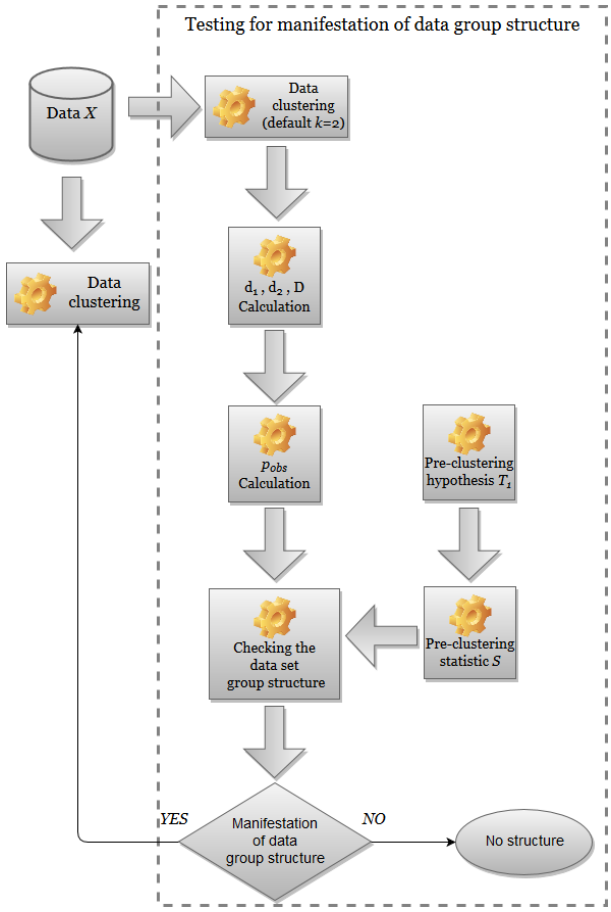

Fig. 3. Data clustering procedure for revealing group structure using the hybrid approach

Example: Let us suppose that we have a two-attribute artificial data set $X = \{x_1, \ldots, x_n\}$ with 100 objects and with normal distribution that is shown in Figure 4. Let us find out if the data set possesses some group structure by testing for revealing the data structure.

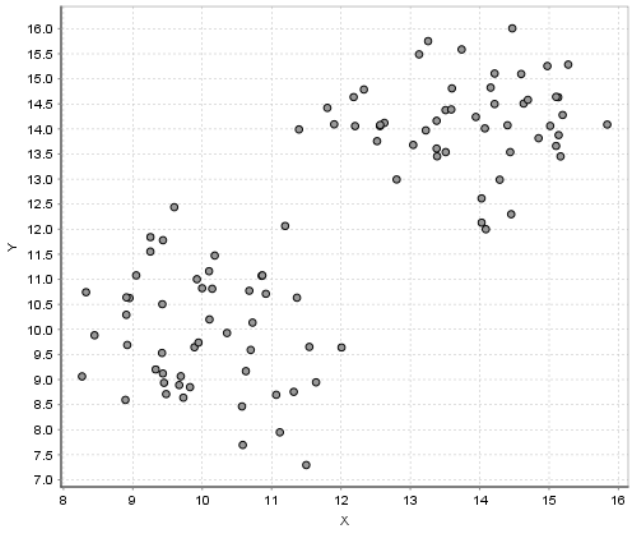

Fig. 4. Artificial data set

Thus, at first, the average pairwise distance D within a data set X is calculated. The obtained value is equal to $D = 3.83$. Then data clustering based on the partitioning method, namely *k*-means algorithm, is to be performed. The number of clusters is set to be two. Then the average pairwise distances d_1 and d_2 within a cluster C_1 and C_2 respectively are calculated. Therefore, $d_1 = 1.88$, and $d_2 = 1.71$. The p_{obs} -value for a given data set is calculated. The obtained value p_{obs} is equal to 0.24 or $p_{obs}^{*} = positive$. Finally, we compare p_{obs} -value with a pre-clustering statistic S and check out the possible revealing of the data group structure. Graphical presentation of the test for revealing data structure is shown in Figure 5.

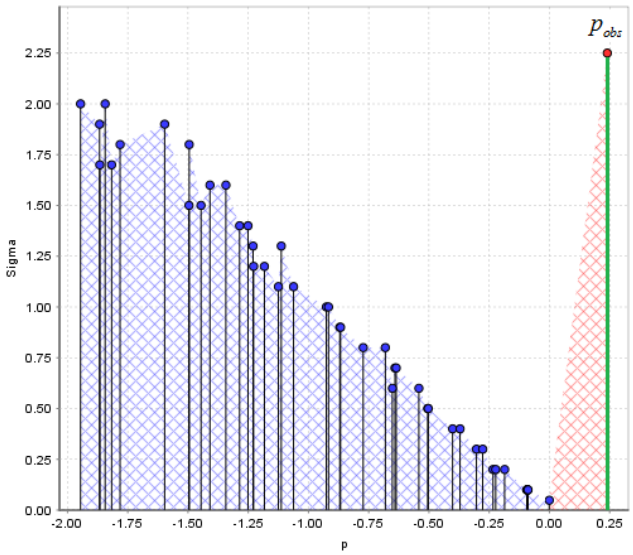

Fig. 5. Graphical presentation of the result of the test for revealing data structure

Blue cross hatched area indicates a pre-clustering statistic and thus the acceptance of pre-clustering hypothesis. However, as Figure 5 shows, the observed p_{obs} value is located in a red cross hatched area and, therefore, $p_{obs} \notin S$ thereby indicates the rejection of pre-clustering hypothesis and, thus, the presence of group structure and a strong justification of further clustering.

4. Experimental results

The test for revealing data structure has been performed for a group of data sets, namely:

Iris data set (150 data objects, 4 attributes, 3 classes), or the Iris flower data set, or Fisher's Iris data set [11] is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis, where each class refers to a type of iris plant with different sepal/petal length/width.

Sonar data set (208 data objects, 61 attributes, 2 classes) is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network [12]. The first 60 objects in the data set represent the energy within a particular frequency band, integrated over a certain period of time. The last column contains the class labels. There are two classes: 0 if the object is a rock and 1 if the object is a mine (metal cylinder).

Ripley data set (250 data objects, 3 attributes, 2 classes) is the nonlinear binary classification problem [13]. The data set consists of two classes, where each class represents a bimodal distribution of input features, which have been generated from two Gaussian mixture distributions with equal covariance.

Artificial *Gaussian Mixture Clusters (GMC)* data set (300 data objects, 2 attributes) is an artificially generated data set with normal distribution and with three widely separated globular form clusters.

Artificial *Single/Double* data set (200 data objects, 2 attributes) is an artificial data set where the number of clusters is questionable and which can be considered as one elongated cluster, as well as two separate clusters that are close to each other.

Artificial *Three Cluster* data set (150 objects, 2 attributes) shows the normal distribution of objects in two-attribute space. The number of clusters is known (k = 3). Clusters are well separated and distributed one above the other.

The test for revealing the data structure with the use of the above presented data sets has been shown in Table 1.

Table 1. Experimental results

Name of data set	p_{obs}^{*}	p_{obs}
Iris	positive	0.27
Sonar	negative	-1.36
Ripley	negative	-0.16
Art. GMC	positive	1.70
Art. Single/Double	negative	-0.47
Atr. Three Cluster	positive	0.91

In the case of *Sonar*, *Ripley* and Artificial *Single/Double* data sets, the acceptance of pre-clustering algorithm is justified, however *Iris*, Artificial *GMC* and Artificial *Three Cluster* data set shows the convincing rejection of a hypothesis.

Most of the fundamental validation tests for revealing the data group structure are based on users' specified thresholds unlike the presented approach. The hybrid test is applied to the real marginal distributions, for example skew or data sets with outliers etc. Moreover, the hybrid approach is not sensitive to changes of implemented data type or other clustering partitioning algorithm. However, a few drawbacks of this approach should also be noted. One of such drawbacks is the dependence on clustering procedure, i.e. on all known drawbacks of partitioning algorithms depending on which we use. This causes the inability to analyze all kinds of data. The critical cases could be artificially created toy data sets informally known as the "half-moons" data sets, where each data point belongs to one of the two "half-moons" or linearly non-separable "ring" or "crater" (the data set consists of a very dense "crater" core with a less dense ring surrounding the core) data sets. Another important drawback is calculating the pre-clustering statistic. If the pre-clustering hypothesis changes a pre-clustering statistic, these changes also cause increase in time and resources for data group revealing. Nevertheless, the subject attracts some theoretical and practical interest and therefore testing for revealing the data group structure remains an integral part of clustering procedure, which cannot be omitted, and, moreover, it can be a viable alternative to the already known validation tests.

References

[1] Mosorov V., Tomczak L.: Image texture defect detection method using fuzzy c-means clustering for visual inspection systems. Arabian Journal for Science and Engineering 39(4)/2014, 3013–3022 [DOI:10.1007/s13369-013-0920-7].

[2] Kumar D., Bezdek J.C., Rajasegarar S., Leckie C., Palaniswami M.: A visual-numeric approach to clustering and anomaly detection for trajectory data. The Visual Computer, December 2015 [DOI:10.1007/s00371-015-1192-x].

[3] Zhang S., Hu W., Wang T., Liu J., Zhang Y.: Speaker Clustering Aided by Visual Dialogue Analysis. Advances in Multimedia Information Processing – PCM 2008. Springer Science + Business Media. 693–702.

[4] Strauss D.J., Riverside C.: A model for clustering. Biometrika 62(2)/ 1975, 467–475 [DOI:10.1093/biomet/62.2.467].

[5] Bock H.H.: On some significance tests in cluster analysis. Journal of Classification 2(1)/1985, 77–108 [DOI:10.1007/bf01908065].

[6] Hartigan J.A., Mohanty S.: The runt test for multimodality. Journal of Classification 9(1)/1992, 63–70 [DOI:10.1007/bf02618468].

[7] Hennig C., Lin C-J.: Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Statistics and Computing 25(4)/2015, 821–833 [DOI:10.1007/s11222-015-9566-5].

[8] Hautaniemi S., Edgren H., Vesanen P. et al.: A novel strategy for microarray quality control using Bayesian networks. Bioinformatics 19(16)/2003, 2031–2038 [DOI:10.1093/bioinformatics/btg275].

[9] Everitt B.S., Landau S., Leese M., Stahl D.: Cluster analysis. John Wiley & Sons, January 7, 2011.

[10] Gordon A.: Studies in Classification, Data Analysis, and Knowledge Organization. Gordon AD. From Data to Knowledge. Springer Science + Business Media 1996, 32–44.

[11] Fisher R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2)/1936, 179–188 [DOI:10.1111/j.1469-1809.1936.tb02137.x].

[12] Gorman R.P., Sejnowski T.J.: Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks 1/1988, 75–89.

[13] Ripley B.D.: Neural networks and related methods for classification. Journal of the Royal Statistical Society - Series B (Methodological) 56(3)/1994, 409–456 [DOI:10.2307/2346118].

D.Sc., Ph.D. Eng. Volodymyr Mosorov
e-mail: w.mosorow@kis.p.lodz.pl

Volodymyr Mosorov received his Ph.D. in 1998 from the State University of Lviv, Ukraine. V.Mosorov was awarded the title of Doctor of Science from AGH University of Science and Technology Krakow Poland in 2009. He is now an associate professor at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data mining and clustering. He has published more than 80 technical articles.

M.Sc. Eng. Taras Panskyi
e-mail: tpanski@kis.p.lodz.pl

Graduated from the Department of Theoretical Radio Engineering and Radio Measurement at Lviv Polytechnic National University, Ukraine. Since 2013, he has been a Ph.D. student at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data clustering, reliability and availability indexes of embedded systems, educational migration.

M.Sc. Eng. Sebastian Biedron
e-mail: sbiedron@wpia.uni.lodz.pl

Graduated from the Department of Science and Mathematics at Lodz University. Since 2012, he has been a court expert at the District Court at the Prague. Since 2013, he has been a Ph.D. student at the Institute of Applied Computer Science of Lodz University of Technology. The supervisor of his Ph.D. thesis is Volodymyr Mosorov.

otrzymano/received:15.06.2016 *przyjęto do druku/accepted: 01.06.2017*