INFORMATICA, 2004, Vol. 15, No. 3, 399–410



# Outlier Detection Based on the Distribution of Distances between Data Points

Vydunas ŠALTENIS
*Institute of Mathematics and Informatics*
*Akademijos 4, 2600 Vilnius, Lithuania*
*e-mail: saltenis@ktl.mii.lt*

Received: January 2004

**Abstract.** A novel approach to outlier detection on the ground of the properties of distribution of distances between multidimensional points is presented. The basic idea is to evaluate the outlier factor for each data point. The factor is used to rank the dataset objects regarding their degree of being an outlier. Selecting the points with the minimal factor values can then identify outliers. The main advantages of the approach are: (1) no parameter choice in outlier detection is necessary; (2) detection is not dependent on clustering algorithms.

To demonstrate the quality of the outlier detection, the experiments were performed on widely used datasets. A comparison with some popular detection methods shows the superiority of our approach.

**Key words:** outlier detection, high-dimensional data, distribution of distances.

## 1. Introduction

The data derived from measurements usually have some amount of noise, which can be viewed as outliers. The outliers also can be viewed as records having abnormal behaviour. The preferable way to deal with outliers is to keep one extra set of outliers, so as not to pollute factual clusters.

For some applications exceptional cases are more interesting than the common ones.

Statistics defines an outlier as a point that does not fit a probability distribution. Data mining has developed its own definitions. The global definition of an outlier introduced in (Knorr *et al*., 1998) considers two parameters $\varepsilon$, $\delta$. A point can be declared an outlier if its $\varepsilon$-neighbourhood contains less than $1 - \delta$ fraction of a whole dataset.

How to describe local outliers? Different subsets of data have different densities and may have different distributions. A point close to a tight cluster can be a more probable outlier than a point that is further away from a more dispersed cluster.

In most of the studies conducted on outlier detection, an object in the dataset may be an outlier or not. For many applications it becomes more meaningful to evaluate a degree of being an outlier for each dataset object. The *local outlier factor* that specifies the degree of outlierness was first introduced and investigated in (Breunig *et al*., 2000). The definition is based on the distance to the $k$-nearest neighbour.

Outliers and clusters in a dataset are related. An outlier means not being in or close to a cluster. According to this approach clustering algorithms generate outliers as by-product (for example, in (Ester *et al.*, 1996)). The data that do not naturally fall into any cluster can be viewed as solitary clusters. However, if a clustering algorithm attempts to find larger clusters, these outliers will be placed in some cluster. This process may create poor clusters.

The outlier factor introduced in this paper also evaluates the degree of being an outlier. The local or global properties of the proposed factor depend on the structure of the dataset.

### 1.1. *Parameter Discussion*

In most outlier detection approaches, the quality of the results greatly depends on a choice of the algorithm parameters. For example, determination of the local outlier factor in (Breunig *et al.*, 2000) can be performed by considering different values of the parameter *MinPts*, specifying a minimum number of objects in the neighborhood of some object. The local outlier factor value changes with a change of the *MinPts* values non-monotonically even for such a pure distribution as the Gaussian distribution.

The notion of *DB*(*pct*, *dmin*) for a global outlier in (Knorr *et al.*, 1998) also contains two parameters: percentage of objects *pct* and minimal distance *dmin*.

Moreover, sometimes the choice of a function in some methods can be arbitrary. For example, the influence function in the density-based clustering method in (Hinneburg *et al.*, 1998) can be an arbitrary function.

Adaptation of the parameter values to the dataset requires additional experimentation. Sometimes the choice of algorithm parameters is based on some a priori recommendations.

Our approach practically does not require the parameter choice or a choice of functions; evaluation of the frequency function in determining the outlier factor can be seen as natural adaptation to the existing structure of distances between data points.

### 1.2. *Distribution of Distances between the Points*

Since visualization of high-dimensional data is rather difficult we want to investigate simple measures to understand the geometry of a given dataset. A useful idea is to analyse the distribution of distances between the points. Brin (1995) draws attention that one way of analysing whether the data may contain clusters is to plot an approximate probability density function of the pairwise distances between all the points in a dataset. If the data contains clusters, then the histogram of distances will show two peaks: a peak representing the distance between points in clusters and that peak representing the average distance between the points. If only one peak is present, then clustering is likely to be difficult.

The distributions of distances between random uniformly chosen vectors in 20 and 50 dimensional hypercubes of side 1 were analysed using $L_1$ and $L_2$ metrics in (Brin, 1995). The author obtained Gaussian-like (not exactly Gaussian) distributions with extremely

narrow peaks. He also stated that correlated or clustered data has somewhat different properties and tends to have a flatter distribution. In his example, taking random blocks from an image, then treating them as 256- or 2500-dimensional vectors and taking the $L_2$ distances between them, he creates a distribution with two maxima. The first maximum near zero distances indicates a great deal of clustering in the data; the second one is common to all multidimensional spaces. We can see in all cases a dominating narrow peak, which is usual for a uniform distribution of data points in a multidimensional hypercube. This peak is different for different space dimensionality.

We will investigate the distribution of distances between random and uniformly chosen vectors of various dimensionality because our outlier detection technique uses these distributions.

The rest of the paper is organized as follows. In the next section, we present theoretical and experimental results relative to the distribution of distances between the uniformly distributed points in a multidimensional hypercube. In Section 3, the basic ideas of evaluation of the outlier factor are introduced. In Section 4, we provide the experimental evaluation of our approach. Section 5 summarizes the results.

## 2. Distribution of Distances between the Uniformly Distributed Points in a Multidimensional Hypercube

Analytical evaluation of the distribution of distances is possible only for a low data dimensionality $n$, therefore a great deal of experimental evaluation has been performed.

### 2.1. *Analytical Evaluation*

Let $n$-dimensional vectors be uniformly distributed in multidimensional unit hypercube $[0..1]^n$. The $L_k$ – norm of a vector $v$ is defined as

$$\|v\|_k = \sqrt[k]{\sum_{i=1}^{n} |v_i|^k}. \tag{1}$$

The $L_k$ – distance between two vectors $v$ and $w$ is

$$\|v - w\|_k. \tag{2}$$

The statistical analysis of distances for the can be found in (Schmitt, 2001).

*In case* $n = 1$, the density function of the distances between two uniformly distributed random values ranging in [0..1] is defined by the derivation of the interval distribution

$$\phi_{|V-V|}(x) = \frac{\mathrm{d}\,(2x - x^2)}{\mathrm{d}\,x} = 2 - 2x \quad \text{for } 0 \leqslant x \leqslant 1 \text{ and } 0 \text{ otherwise.} \tag{3}$$

The expecting value and the standard deviation of the distances therefore are:

$$E_{|V-V|} = \frac{1}{3} \tag{4}$$

and

$$\sigma_{|V-V|} = \frac{1}{\sqrt{18}}. \tag{5}$$

*In case* $n > 1$, we may find only approximate evaluations in (Schmitt, 2001) that the expecting value converges in probability to the constant

$$E_{|V-V|} \underset{p}{\longrightarrow} \sqrt{\frac{n}{6}}. \tag{6}$$

Experiments with higher dimensions confirm this result. For a low number of dimensions $n$ the expecting value evaluation (6) differs (especially if $n = 1$, the correct value is 1/3, instead).

The analytical evaluation of standard deviation $\sigma_{|V-V|}$ is very hard in case $n > 1$. Experiments with a high number of dimensions show that the standard deviation converges towards the constant 0.24.

### 2.2. *Experimental Evaluation*

In a hypercube of side length 1 two points are selected from a uniform distribution, and their Euclidean distance $L_2$ is computed. This is repeated for $1\,000\,000$ times. The results are the computed relative frequency distribution $f^u(d)$ of distances between the points, normalized to the total sum 1. They are illustrated in Fig. 1 for dimensionalities 1–7, and their values are presented in table in the Appendix.

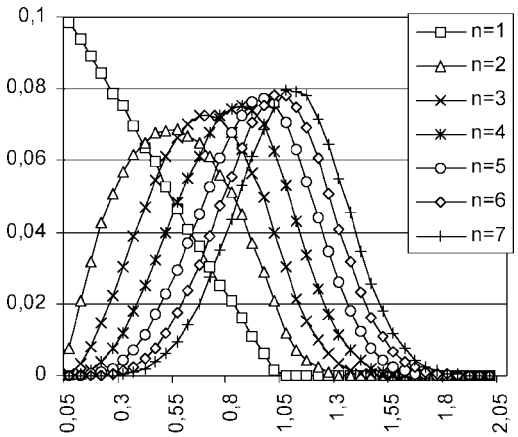


Fig. 1. Computed relative frequency distribution of $L_2$ distances between the uniformly distributed multidimensional points for dimensionalities $n = 1 - 7$.

The frequency distributions of distances of uniformly distributed multidimensional points are extremely non-uniform, especially for higher dimensions. This fact of dominating narrow peak existence is a disadvantage if we want to use these distributions in the analysis of multidimensional data with a view to extract some properties on point outlierness.

## 3. The Basic Idea and the Algorithm

The basic idea is to eliminate the influence of a dominating narrow peak in the relative frequency distribution, and analyse only the difference between the frequency distribution of a given dataset $f^n(d)$ and the corresponding distribution of uniformly distributed points $f^u(d)$ of the same dimensionality. The difference function

$$f(d) = f^n(d) - f^u(d) \tag{7}$$

may be treated as a *frequency function*.

If the points in the given dataset are uniformly distributed, we obtain frequency function with the values near to zero for all interpoint distances.

If the points are not uniformly distributed, the greatest values of frequency function will indicate the most frequent and most characteristic interpoint distances; the lowest values will indicate rare, atypical distances.

The frequency function is similar to the influence function introduced in (Hinneburg *et al*., 1998) which can be an arbitrary (for example, Square Wave or Gaussian) function.

An example of the frequency function for a two-dimensional dataset is illustrated in Fig. 2. The dataset consists of two normally distributed data clusters; the distance between them is equal to 0.8. The function has two peaks typical for data with clusters.

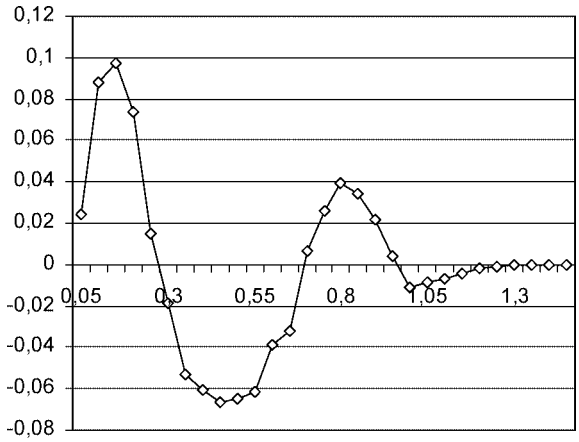


Fig. 2. An example of the frequency function for a two-dimensional dataset with two clusters.

For each data point $i$ the *outlier factor*

$$R_i = 1/m \sum_{\substack{j=1 \\ j \neq i}}^{m} f\big(d(X_i, X_j)\big) \tag{8}$$

can be calculated. Here $m$ is the number of data objects, $d(X_i, X_j)$ is the distance between the $i$-th and $j$-th points, and $f(d)$ is the difference function (7).

It is obvious that the values of *outlier factor* are limited between $-1$ and $1$.

The outliers will have the lowest values of the outlier factor $R$, because the distances between the outlier points and the rest data points will be atypical and the corresponding frequency function values in (7) will be low.

**The algorithm** naturally follows from the basic idea. It works in four steps:

*Step* 1: The frequency distribution $f^n(d)$ of a given dataset is evaluated.

*Step* 2: The *frequency function* is calculated according to (7) and using the values of the table in Appendix.

*Step* 3: For each data point the *outlier factor R* is calculated according to (8).

*Step* 4: The data points with the lowest values of the *outlier factor* are treated as outliers.

Obviously the algorithm's complexity is $\mathrm{O}(m^2)$, where $m$ is the number of data objects. However the estimation of the frequency function and outlier factor may be based only on a part of data points. Such an approximation may improve the complexity, but its influence requires further investigations.

## 4. Tests and Results

We discuss the results from the test datasets, used to evaluate the performance of the proposed approach to the outlier detection. The outlier detection methods used for comparison are selected on the base of the availability of implementation.

The four methods used for comparison are:

- Donoho–Stahel (see (Knorr *et al.*, 2001)),
- Hadi (see (Hadi, 1994)),
- MML clustering (see (Oliver *et al.*, 1996)), and
- Replicator neural networks (RNN) (see (Williams *et al.*, 2002)).

Other detection methods, for example, Atkinson (1994), Billor *et al.* (2000), deBoer *et al.* (2000), Knorr *et al.* (2000), Kollios *et al.* (2001), are related to the four included methods.

### 4.1. *HBK Dataset*

The HBK dataset is an artificially constructed 4-dimensional dataset with 75 data points including 14 outliers (Hawkins *et al.*, 1984). Regression approaches can find only the first

10 outliers, but cannot find scattered data points 11–14 ("true" outliers) although they lie far away from the main data cluster.

The results of the proposed outlier detection method and the four methods used for comparison on the HBK dataset are summarized in Table 1. The values in columns 1–4 of Table 1 are taken from (Williams *et al*., 2002).

Donoho–Stahel and Hadi methods rank all 14 outliers at the top 14 places but the "true" outliers are not at the top 1–4 places.

The MML method identifies only four scattered "true" outliers but the rest are placed not adequately.

The RNN method has 14 outliers at the top 16 places and has located all the "true" outliers in a single cluster at the 1–4 places.

The proposed method demonstrates the best detection of all and "true" outliers.

### 4.2. *Wood Dataset*

The four outliers (data points 4, 6, 8, and 19) in the Wood dataset (Draper *et al*., 1966) are not easily identifiable according (Rousseeuw *et al*., 1987). The number of data is equal

Table 1

Top 20 outliers for Donoho–Stahel, Hadi, MML Clustering, RNN and the proposed method on the HBK dataset. The first 14 outliers are written in Italics. The four scattered "true" outliers (11, 12, 13, 14) lie far away from the main data cluster; they are underlined

| Dono–Stahel | Hadi | MML Clustering | RNN | Proposed method |
|---|---|---|---|---|
| *14* | *14* | *12* | *14* | *11* |
| *4* | *4* | *14* | *12* | *14* |
| *3* | *5* | *13* | *13* | *13* |
| *5* | *3* | *11* | *11* | *12* |
| *9* | *9* | *4* | *7* | *5* |
| *7* | *7* | 53 | *6* | *3* |
| *10* | *12* | *7* | *8* | *1* |
| *6* | *10* | 47 | *3* | *7* |
| *2* | *6* | 68 | *1* | *8* |
| *8* | *2* | 62 | *2* | *10* |
| *12* | *8* | 60 | *10* | *2* |
| *13* | *13* | 34 | *5* | *9* |
| *1* | *11* | 43 | 16 | *6* |
| *11* | *1* | 27 | 49 | *4* |
| 53 | 53 | 49 | *4* | 15 |
| 47 | 47 | 16 | *9* | 27 |
| 68 | 68 | *1* | 20 | 41 |
| 75 | 43 | 38 | 51 | 49 |
| 43 | 34 | 20 | 21 | 20 |
| 70 | 60 | 61 | 31 | 35 |

to 20, and the dimensionality is equal to 6.

The results of the proposed outlier detection method and the four methods used for comparison on the Wood dataset are summarized in Table 2. The values in columns 1–4 of Table 2 are taken from (Williams *et al.*, 2002).

Hadi, MML Clustering and RNN methods did not successfully identify all outliers. According to the MML method the outliers are identified only in the last 4 places.

The proposed and Donoho–Stahel methods demonstrate the best outlier detection. They rank all 4 outliers successfully, at the top 4 places. The order of all the 4 outliers is in fact the same for both methods.

In Fig. 3 we can see that the ranked values of the outlier factor sharply distinguish four "real" outliers: ten "not true" outliers in HBK data, and similarly four outliers in Wood data from the remainder of the points.

Using test datasets, we demonstrate that the proposed outlier factor can be used to find outliers that cannot be otherwise identified by the existing approaches.

Table 2

Top 20 outliers for Donoho–Stahel, Hadi, MML Clustering, RNN and the proposed method on the Wood dataset. 4 outliers (points 4, 6, 8, and 19) are "true"; they are in Italics and underlined

| Dono–Stahel | Hadi | MML Clustering | RNN | Proposed method |
|---|---|---|---|---|
| *19* | 7 | 10 | 13 | *19* |
| *6* | 11 | 12 | 10 | *8* |
| *8* | 16 | 20 | *6* | *6* |
| *4* | 9 | 11 | 9 | *4* |
| 11 | 12 | 7 | 20 | 2 |
| 7 | 10 | 13 | 12 | 12 |
| 16 | *19* | 1 | *19* | 14 |
| 5 | 17 | 9 | 7 | 10 |
| 12 | 18 | 17 | *4* | 13 |
| 14 | 3 | 18 | 18 | 15 |
| 10 | 20 | 3 | 16 | 11 |
| 9 | 1 | 16 | 11 | 16 |
| 1 | 13 | 2 | 1 | 17 |
| 3 | *8* | 14 | 5 | 3 |
| 18 | 5 | 5 | 15 | 5 |
| 17 | *6* | 15 | 17 | 7 |
| 20 | 14 | *4* | 14 | 1 |
| 15 | *4* | *19* | 2 | 20 |
| 13 | 15 | *8* | *8* | 18 |
| 2 | 2 | *6* | 3 | 9 |

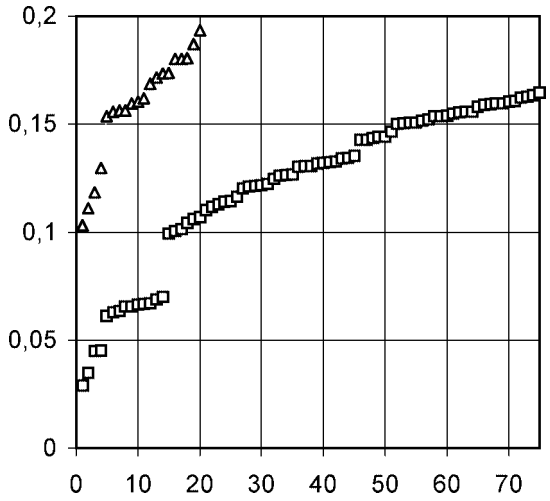


Fig. 3. Ranked values of the outlier factor for all the points of HBK dataset (*square data points*) and Wood dataset (*triangular data points*).

## 5. Conclusions

In this paper, we propose a new approach for the multivariate outlier detection.

The outlier factor notion is based on the frequency distribution of distances between the multidimensional points. It naturally evaluates the degree of being an outlier.

Our approach does not require the parameter choice or a choice of functions; evaluation of the frequency function in determining the outlier factor can be seen as a natural adaptation to the existing structure of distances between data points.

The comparison with some popular outlier detection methods shows the superiority of our approach.

There are two directions in ongoing research. The first one is to improve and investigate the performance of computation. The second one is to expand the capabilities of the proposed approach to clustering data containing large amounts of noise.

## 6. Acknowledgements

This research was partially supported by the Lithuanian State Science and Studies Foundation, Grant No. C 03013.

**Appendix. Relative Frequency Distribution of Distances between the Points for Dimensionalities 2–7**

| Intervals of distances | $n = 2$ | $n = 3$ | $n = 4$ | $n = 5$ | $n = 6$ | $n = 7$ |
|---|---|---|---|---|---|---|
| 0 – 0.1 | 0.028248 | 0.003724 | 0.000438 | 6.21E-05 | 4E-06 | |
| 0.1 – 0.2 | 0.074865 | 0.023037 | 0.005431 | 0.001139 | 0.000228 | 3.8E-05 |
| 0.2 – 0.3 | 0.107724 | 0.052889 | 0.019487 | 0.006392 | 0.001928 | 0.000515 |
| 0.3 – 0.4 | 0.126370 | 0.085838 | 0.043165 | 0.018647 | 0.007465 | 0.002863 |
| 0.4 – 0.5 | 0.136276 | 0.115526 | 0.072224 | 0.039776 | 0.019664 | 0.009668 |
| 0.5 – 0.6 | 0.135251 | 0.136354 | 0.103483 | 0.066080 | 0.041315 | 0.023417 |
| 0.6 – 0.7 | 0.126206 | 0.144947 | 0.128833 | 0.097856 | 0.069383 | 0.047321 |
| 0.7 – 0.8 | 0.107135 | 0.139666 | 0.146991 | 0.128757 | 0.102671 | 0.078755 |
| 0.8 – 0.9 | 0.082220 | 0.119686 | 0.149317 | 0.148947 | 0.133778 | 0.114418 |
| 0.9 – 1.0 | 0.049624 | 0.089606 | 0.132659 | 0.152905 | 0.153391 | 0.143333 |
| 1.0 – 1.1 | 0.018749 | 0.051632 | 0.096332 | 0.133724 | 0.153213 | 0.158745 |
| 1.1 – 1.2 | 0.006344 | 0.024941 | 0.057946 | 0.097756 | 0.129481 | 0.149526 |
| 1.2 – 1.3 | 0.000961 | 0.009399 | 0.028386 | 0.060561 | 0.091285 | 0.119391 |
| 1.3 – 1.4 | 2.6E-05 | 0.002358 | 0.011279 | 0.030458 | 0.054749 | 0.079762 |
| 1.4 – 1.5 | | 0.000350 | 0.003267 | 0.012208 | 0.026669 | 0.043792 |
| 1.5 – 1.6 | | 4.8E-05 | 0.000649 | 0.003700 | 0.010444 | 0.019548 |
| 1.6 – 1.7 | | | 9.81E-05 | 0.000887 | 0.003315 | 0.006525 |
| 1.7 – 1.8 | | | 8.01E-06 | 0.000134 | 0.000807 | 0.001880 |
| 1.8 – 1.9 | | | 4E-06 | 1.2E-05 | 0.000186 | 0.000434 |
| 1.9 – 2.0 | | | | | 2.2E-05 | 5.81E-05 |

## References


Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, **89**, 1329–1339.

Billor, N., A.S. Hadi and P.F. Velleman (2000). BACON: Blocked adaptive computationally-efficient outlier nominators. *Computational Statistics and Data Analysis*, **34**, 279–298.

deBoer, P., and V. Feltkamp (2000). Robust multivariate outlier detection. *Technical Report 2*, Statistics Netherlands, Department. of Statistical Methods.
`http://neon.vb.cbs.nl/rsm/P_su2000.htm.`

Breunig, M.M., H.-P. Kriegel, R.T. Ng and J. Sander (2000). LOF: identifying density-based local outliers. In *Proceedings of the ACM SIGMOD Conference*. Dallas, TX. pp. 93–104.

Brin, S. (1995). Near neighbor search in large metric spaces. In *Proceedings of the 21st International Conference on Very Large Databases* (VLDB-1995). Morgan Kaufmann, Zurich, Switzerland. pp. 574–584.

Draper, N.R., and H. Smith (1966). *Applied Regression Analysis*. John Wiley and Sons, New York.

Ester, M., H. Kriegel, J. Sander and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. pp. 226–231.

Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society*, B, **56**(2), 393–396.

Hawkins, D.M., D. Bradu and G.V. Kass (1984). Kass location of several outliers in multiple regression data using elemental sets. *Technometrics*, **26**, 197–208.

Hinneburg, A., and D. Keim (1998). An efficient approach to clustering large multimedia databases with noise. In *Proceedings of the 4th ACM SIGKDD*. New York, NY. pp. 58–65.

Knorr, E.M., and R.T. Ng (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference On Very Large Data Bases*. New York, NY. pp. 392–403.

Knorr, E., R. Ng and V. Tucakov (2000). Distance-based outliers: Algorithms and applications. *Very Large Data Bases*, **8**(3–4), 237–253.

Knorr, E.M., R.T. Ng and R.H. Zamar (2001). Robust space transformations for distance-based operations. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD01)*. San Francisco, California. pp. 126–135.

Kollios, G., D. Gunopoulos, N. Koudas and S. Berchtold (2001). An efficient approximation scheme for data mining tasks. In *Proceedings of the International Conference on Data Engineering (ICDE01)*. pp. 453–463.

Oliver, J.J., R.A. Baxter and C.S. Wallace (1996). Unsupervised Learning using MML. In *Proceedings of the Thirteenth International Conference* (ICML 96). Morgan Kaufmann Publishers, San Francisco. pp. 364–372.

Rousseeuw, P.J., and A.M. Leroy (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

Schmitt, I. (2001). *Nearest Neighbor Search in High Dimensional Space by Using Convex Hulls*. Preprint No. 6/01, Fakultat fur Informatik, Universitat Magdeburg, 1–30.

Williams, G., R. Baxter, H. He, S. Hawkins and L. Gu (2002). *A Comparative Study of Replicator Neural Networks for Outlier Detection in Data Mining*. CSIRO Technical Report CMIS-02/102, Canbera, Australia, 1–16.


**V. Šaltenis** graduated from the Kaunas Technological Institute, Lithuania. He received PhD degree from the Moscow Energy Institute of the USSR Academy of Sciences in 1966 and the degree of habil. dr. from the Institute of Mathematics and Informatics, Vilnius in 1998. He is a principal researcher of the Systems Analysis department at the Institute of Mathematics and Informatics, Lithuania. His present research interests include data analysis, theory and applications of the structure of optimisation problems, multicriteria decision support systems, unconventional computing.

## Išsiskiriančių duomenų taškų paieška, grindžiama taškų tarpusavio atstumais

Vydūnas ŠALTENIS

Straipsnyje įvestas duomenų taškų išskirtinumo matas, leidžiantis naujai ir efektyviai vykdyti daugiamačių išsiskiriančių duomenų taškų paiešką. Matas grindžiamas taškų tarpusavio atstumų analize. Tam lyginamas atstumų tarp duomenų taškų pasiskirstymas su daugiamačiame kube tolygiai pasiskirsčiusių taškų tarpusavio atstumų pasiskirstymu. Pasiūlytiems algoritmams nebūtinas parametrų parinkimas – metodas natūraliai prisitaiko prie analizuojamų duomenų struktūros.

Eksperimentai pasiūlytu algoritmu analizuojant plačiai paplitusius testinius išsiskiriančių duomenų taškų paieškos uždavinius parodė jo geresnį veikimą lyginant su žinomais metodais.