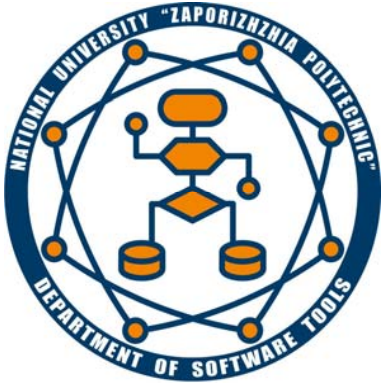
ViMaCs
Virtual Master Cooperation
Data Science


**Ministry of Education and Science of Ukraine**
**"Zaporizhzhia Polytechnic"**
**National University**

## Lecture

# "THE DATA HASHING TRANSFORMATIONS FOR DIMENSIONALITY REDUCTION IN DIAGNOSTIC AND RECOGNITION PROBLEMS"

**Sergey A. Subbotin,**
Professor, Dr. hab. Sc.

2021

# THE IMPORTANCE OF COMPUTATIONAL DIAGNOSIS

| | |
|---|---|
| **Sphere:** decision making in technology and medicine, where the expert knowledge is missing or insufficient.<br><br>**Theoretical basis:** pattern recognition, machine learning.<br><br>**Decision basis:** precedents (instances, examples, exemplars, cases). | **A general problem is** a need of ***model complexity reduction*** and a ***model construction speed increasing***.<br>It is caused by that the samples frequently characterized by a big number of features or contain a big number of instances (precedents).<br><br>To reduce the model complexity and to speed-up it's training process we need to ***reduce the data dimensionality***. |

# THE DATA DIMENSIONALITY REDUCTION

The ***data dimensionality reduction*** can be performed by

| | |
|---|---|
| – a ***feature selection***.<br>It needs, as a rule, an exhaustive search of possible feature combination with the least number of features that provides an acceptable accuracy. When the number of features and instances are big such technique require a lot of time. | – a ***sample selection***.<br>Unfortunately, it involves necessity of computation and using a distance matrix in *N*-dimensional feature space that causes a combinatorial complexity of large data processing. |

The ***way to solve these problems*** is the using of ***transformation from the multidimensional space of initial features to a one-dimensional axis*** for data dimensionality reduction.

# THE DATA DIMENSIONALITY REDUCTION

The known popular **methods of transformation for data dimensionality reduction:**

– **Principal Component Analysis (PCA)**

– **Semidefinite Embedding (SDE)**

– **Multifactor Dimensionality Reduction (MDR)**

– **Non-linear Dimensionality Reduction (NLDR)**

– **Partial Least Squares (PLS)**

– **Independent Component Analysis (ICA)**

– **Vasil'yev theory of reduction**

– **Babak generalized variable method**

**The common disadvantages of known methods**

The known methods **require the calculation of distances** between instances or feature correlation coefficients and for a large-scale problem they are hardly applicable in practice due to big requirements of time and computer memory in the process of determining the transformation parameters and in the process of transformation execution.

This situation is additionally compounded by that the **number of known transformations and their modifications is very big** and there are no any formal criteria to analyze their quality, as well as to **select the best** available **transformation** for a particular task.

# THE PROBLEM STATEMENT

Suppose we have an initial (original) **sample** $X = <x, y>$ the set of $S$ precedents describing dependence $y(x)$, $x = \{x^s\}$, $y = \{y^s\}$, $s = 1, 2, ..., S$, characterized by a set of $N$ input features $\{x_j\}$, $j = 1, 2, ..., N$, where $j$ is the number of feature, and an output feature $y$.

Each $s$-th **precedent** can be represented as

$$<x^s, y^s>, x^s = \{x^s_j\},$$

where

$x^s_j$ is the value of $j$-th input feature,

$y^s$ is the value of output feature for the $s$-th precedent (instance) of the sample, $y^s \in \{1, 2, ..., K\}$,

$K$ is the number of classes, $K > 1$.

Then ***the problem of the sample X dimensionality reduction by creation of artificial feature*** can be formally represented as follows:

**find a transformation** $H: X \to I$,

which for each instance $x^s = \{x_j\}$ determine the coordinate $I^s$ on the generalized axis $I$ and thus provides a mapping of instances of different classes to the different intervals of the generalized axis.

Since, as a rule, known transformations do not guarantee an exact solution of this problem, further problem arises of designing of indicators to quantify the quality of the transformation and to compare the results of the various transformations between themselves to choose the best transformation of the set.

## SIMPLE PAIRWISE FEATURE TRANSFORMATIONS

Let we have two **original input features** $x_i$ and $x_j$.

We want to **define such transformation to generalized axis** (artificially constructed feature) $x_*^s = \aleph_z(x_i^s, x_j^s)$, that will provide better class separation comparing both original features.

**Transformations:**

$$\aleph_{\mathrm{I}}(x_i^s, x_j^s) = x_i^s + x_j^s;$$

$$\aleph_{\mathrm{II}}(x_i^s, x_j^s) = x_i^s x_j^s;$$

$$\aleph_{\mathrm{III}}(x_i^s, x_j^s) = {x_i^s}^2 + {x_j^s}^2;$$

$$\aleph_{\mathrm{IV}}(x_i^s, x_j^s) = \frac{\widetilde{C}_i^1\widetilde{C}_i^2 + \widetilde{C}_j^1\widetilde{C}_j^2 + \left(\widetilde{C}_i^1 - \widetilde{C}_i^2\right)\widetilde{x}_i^s + \left(\widetilde{C}_j^1 - \widetilde{C}_j^2\right)\widetilde{x}_j^s}{\sqrt{\left(\widetilde{C}_i^1 - \widetilde{C}_i^2\right)^2 + \left(\widetilde{C}_j^1 - \widetilde{C}_j^2\right)^2}},$$

where $\widetilde{C}_i^1, \widetilde{C}_i^2, \widetilde{C}_j^1, \widetilde{C}_j^2, \widetilde{x}_i^s, \widetilde{x}_j^s$ are defined by the following method.

0. Compute coordinates of etalon for each class:

$$C_i^q = \frac{1}{S^q}\sum_{s=1}^{S}\left\{x_i^s \middle| y^s = q\right\}, i = 1, 2, ..., N; q = 1, 2,$$

where $q$ is a class number, $S^q$ is a number of instances in a sample, belonging to $q$-th class, $C^q{}_i$ is a coordinate of etalon of $q$-th class by the axis of $i$-th feature.

1. Set: $\widetilde{C}_i^1 = C_i^1, \widetilde{C}_i^2 = C_i^2, \widetilde{C}_j^1 = C_j^1, \widetilde{C}_j^2 = C_j^2, \widetilde{x}_i^s = x_i^s, \widetilde{x}_j^s = x_j^s.$

2. If $\left(\widetilde{C}_i^1 = \widetilde{C}_i^2\right) \wedge \left(\widetilde{C}_j^1 = \widetilde{C}_j^2\right)$ then execute steps 2.1–2.2, otherwise go to step 3.

2.1. If $\widetilde{C}_i^1 = \widetilde{C}_i^2 = \widetilde{C}_j^1 = \widetilde{C}_j^2 = 0$ then set: $\widetilde{C}_i^2 = 1, \widetilde{C}_j^2 = 1.$

2.2. Set:: $\widetilde{C}_i^1 = 0, \widetilde{C}_j^1 = 0.$

3. If $\left(\widetilde{C}_i^1\right)^2 + \left(\widetilde{C}_j^1\right)^2 < \left(\widetilde{C}_i^2\right)^2 + \left(\widetilde{C}_j^2\right)^2$ then set: $t_i = \widetilde{C}_i^1, t_j = \widetilde{C}_j^1, \widetilde{C}_i^1 = \widetilde{C}_i^2, \widetilde{C}_j^1 = \widetilde{C}_j^2, \widetilde{C}_i^2 = t_i, \widetilde{C}_j^2 = t_j.$

$$\aleph_V(x_i^s, x_j^s) = \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2;$$

where $\overline{C}_i = \frac{C_i^1 + C_i^2}{2}, i = 1,2,..., N;$

Normalized angle between the line connecting projected point with point laying in the middle of the segment, connecting class centers (etalons), and и line running through the point laying in the middle of segment, connecting class centers, parallel to *i*-th feature axis:

$$\aleph_{\mathrm{VI}}(x_i^s, x_j^s) = \begin{cases} \dfrac{\gamma}{2\pi}, x_i^s - \overline{C}_i > 0, x_j^s - \overline{C}_j > 0, \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2 > 0; \\ \dfrac{\gamma + \dfrac{\pi}{2}}{2\pi}, x_i^s - \overline{C}_i < 0, x_j^s - \overline{C}_j > 0, \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2 > 0; \\ \dfrac{\gamma + \pi}{2\pi}, x_i^s - \overline{C}_i \le 0, x_j^s - \overline{C}_j \le 0, \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2 > 0; \\ \dfrac{\gamma + \dfrac{3\pi}{2}}{2\pi}, x_i^s - \overline{C}_i \ge 0, x_j^s - \overline{C}_j > 0, \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2 > 0; \\ 0, \left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2 \le 0, \end{cases}$$

where

$$\gamma = \arcsin\left( \frac{x_j^s - \overline{C}_j}{\sqrt{\left(\overline{C}_i - x_i^s\right)^2 + \left(\overline{C}_j - x_j^s\right)^2}} \right).$$

Combinations of angle and distance:

$$\aleph_{\mathrm{VII}}(x_i^s, x_j^s) = \aleph_{\mathrm{V}}(x_i^s, x_j^s)\aleph_{\mathrm{VI}}(x_i^s, x_j^s);$$

$$\aleph_{\mathrm{VIII}}(x_i^s, x_j^s) = \frac{\aleph_{\mathrm{VI}}(x_i^s, x_j^s)}{\aleph_{\mathrm{V}}(x_i^s, x_j^s)};$$

$$\aleph_{\mathrm{IX}}(x_i^s, x_j^s) = \aleph_{\mathrm{V}}(x_i^s, x_j^s)\cos\left(\aleph_{\mathrm{VI}}(x_i^s, x_j^s)\right);$$

$$\aleph_{\mathrm{X}}(x_i^s, x_j^s) = \aleph_{\mathrm{V}}(x_i^s, x_j^s)\Big|_{\overline{C}=C^1}; \aleph_{\mathrm{XI}}(x_i^s, x_j^s) = \aleph_{\mathrm{VI}}(x_i^s, x_j^s)\Big|_{\overline{C}=C^1};$$

$$\aleph_{\mathrm{XII}}(x_i^s, x_j^s) = \aleph_{\mathrm{V}}(x_i^s, x_j^s)\Big|_{\overline{C}=C^2}; \aleph_{\mathrm{XIII}}(x_i^s, x_j^s) = \aleph_{\mathrm{VI}}(x_i^s, x_j^s)\Big|_{\overline{C}=C^2};$$

$$\aleph_{\mathrm{XIV}}(x_i^s, x_j^s) = \frac{\aleph_{\mathrm{X}}(x_i^s, x_j^s)}{\aleph_{\mathrm{XII}}(x_i^s, x_j^s)}; \aleph_{\mathrm{XV}}(x_i^s, x_j^s) = \frac{\aleph_{\mathrm{XI}}(x_i^s, x_j^s)}{\aleph_{\mathrm{XIII}}(x_i^s, x_j^s)};$$

$$\aleph_{\mathrm{XVI}}(x_i^s, x_j^s) = \aleph_{\mathrm{VI}}(x_i^s, x_j^s)\Big|_{\overline{C}=\widehat{C}},$$

where

$$\widehat{C}_i^q = \frac{1}{2}\left(\max_{s=1,2,\ldots,S}\left\{x_i^s \middle| y^s = q\right\} + \min_{s=1,2,\ldots,S}\left\{x_i^s \middle| y^s = q\right\}\right), i = 1, 2, \ldots, N; q = 1, 2.$$

# INDICATORS OF FEATURE TRANSFORMATION PROPERTIES

**Absolute individual estimation** of informativity of *j*-th feature relative to class number:

$$I_j = \frac{K}{N_j},$$

where $K$ is a number of classes in a sample, $N_j$ is a number of intervals of classes on the axis of *j*-th feature.

The bigger the $I_j$ value the better the *j*-th feature.

**Relative individual estimation** of informativity of *j*-th feature:

$$I_j' = \frac{\min\limits_{j=1,2,..,N}\{N_j\}}{N_j}.$$

**Fechner correlation coefficient:**

$$r_j^{\Phi.} = \frac{\left| 2\sum_{s=1}^{S}\left\{1\middle|\operatorname{sign}(x_j^s - \bar{x}_j) = \operatorname{sign}(y^s - \bar{y})\right\} - S \right|}{\left| 2\sum_{s=1}^{S}\left\{1\middle|\operatorname{sign}(x_j^s - \bar{x}_j) = \operatorname{sign}(y^s - \bar{y})\right\} + S \right|},$$

where

$$\bar{x}_j = \frac{1}{S}\sum_{s=1}^{S} x_j^s,$$

$$\bar{y} = \frac{1}{S}\sum_{s=1}^{S} y^s,$$

$$\operatorname{sign}(a) = \begin{cases} 1, a > 0; \\ 0, a = 0; \\ -1, a < 0. \end{cases}$$

The bigger the Fechner correlation coefficient value the better the *j*-th feature

## INDICATORS OF FEATURE TRANSFORMATION PROPERTIES

### Module of sign correlation coefficient:

$$r_j^{\text{зн.}} = \left| \frac{\frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(x_j^s - \bar{x}_j) > 0, \operatorname{sign}(y^s - \bar{y}) > 0\right.\right\} - \left(\frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(x_j^s - \bar{x}_j) > 0)\right.\right\}\right)\left(\frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(y^s - \bar{y}) > 0\right.\right\}\right)}{\sqrt{\left(\frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(x_j^s - \bar{x}_j) > 0)\right.\right\}\right)\left(\frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(y^s - \bar{y}) > 0\right.\right\}\right)\left(1 - \frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(x_j^s - \bar{x}_j) > 0)\right.\right\}\right)\left(1 - \frac{1}{S}\sum_{s=1}^{S}\left\{1\left|\operatorname{sign}(y^s - \bar{y}) > 0\right.\right\}\right)}} \right|.$$

The bigger the module of sign correlation coefficient value the better the *j*-th feature.

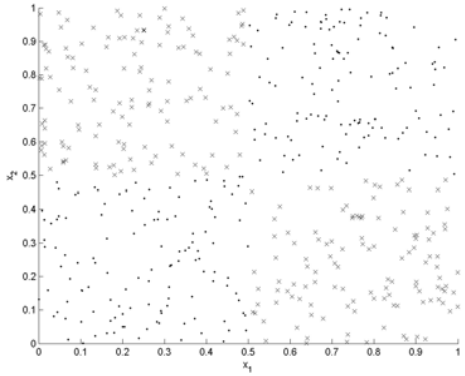
### Module of pair correlation coefficient:

$$r_j = \left| \frac{\sum_{s=1}^{S}\left(x_j^s - \bar{x}_j\right)\left(y^s - y^s\right)}{\sqrt{\sum_{s=1}^{S}\left(x_j^s - \bar{x}_j\right)^2 \sum_{s=1}^{S}\left(y^s - \bar{y}\right)^2}} \right|.$$

The bigger the module of correlation coefficient value the better the *j*-th feature.

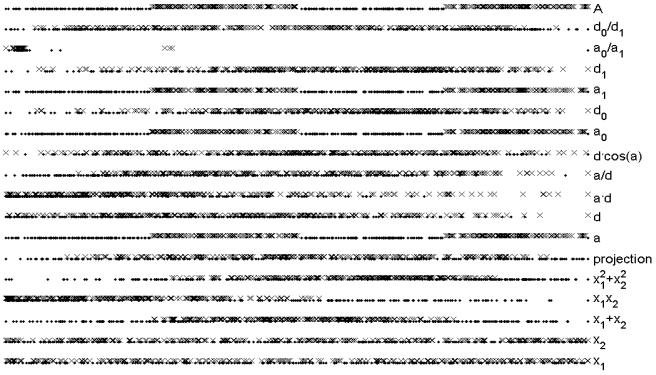
# EXPERIMENTS ON PAIRWISE FEATURE TRANSFORMATION

## Source sample in two features coordinates

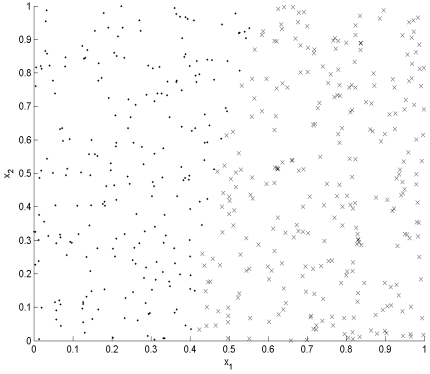



a

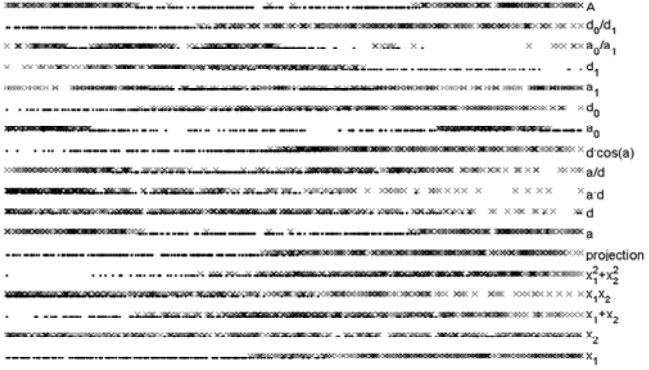
## Transformed coordinates




b



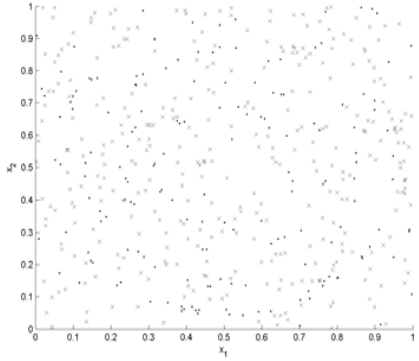

c



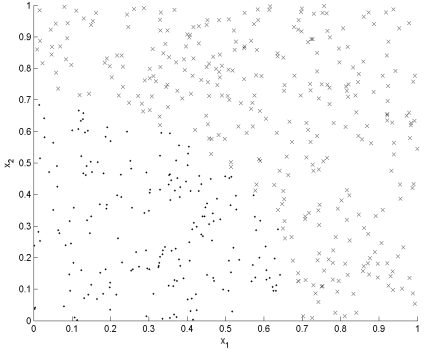

d

# EXPERIMENTS ON PAIRWISE FEATURE TRANSFORMATION

## Source sample in two features coordinates




e




g

## Transformed coordinates


f


h

Table 1 – Results for transformations to generalized axis for problems with class interpenetration

| $\aleph_z(x_i^s, x_j^s)$ | Designations at figures | Averaged estimations of criterions | | | | |
|---|---|---|---|---|---|---|
| | | $I_z$ | $I_z'$ | $r_z^{зн.}$ | $r_z^{Ф.}$ | $r_z$ |
| $x_1$ | $x_1$ | 0.048062 | 0.40514 | 0.083567 | 0.081409 | 0.083566 |
| $x_2$ | $x_2$ | 0.048 | 0.40282 | 0.082538 | 0.080053 | 0.08225 |
| I | $x_1+x_2$ | 0.075671 | 0.45551 | 0.099877 | 0.097898 | 0.082757 |
| II | $x_1x_2$ | 0.06191 | 0.43522 | **0.1407** | **0.14476** | 0.17913 |
| III | $x_1^2+x_2^2$ | 0.067894 | 0.44522 | 0.094437 | 0.093204 | 0.082505 |
| IV | projection | 0.060936 | 0.43241 | 0.13243 | 0.12862 | 0.1299 |
| V | d | **0.24188** | **0.71953** | 0.0824 | 0.081198 | **0.23469** |
| VI | a | 0.052236 | 0.45301 | 0.087407 | 0.084565 | 0.084465 |
| VII | a·d | 0.051362 | 0.41971 | **0.14753** | **0.15017** | 0.1717 |
| VIII | a/d | 0.054178 | 0.41434 | 0.14027 | 0.13875 | 0.21601 |
| IX | d·cos(a) | 0.048576 | 0.4063 | 0.077516 | 0.073505 | 0.068288 |
| X | $d_0$ | 0.2091 | 0.66548 | 0.078479 | 0.081677 | **0.2277** |
| XI | $a_0$ | 0.051011 | 0.4369 | 0.094765 | 0.093334 | 0.094446 |
| XII | $d_1$ | **0.22226** | **0.68901** | 0.084798 | 0.079413 | 0.22734 |
| XIII | $a_1$ | 0.051901 | 0.44949 | 0.088465 | 0.086031 | 0.088954 |
| XIV | $d_0/d_1$ | 0.066816 | 0.43964 | 0.11756 | **0.15806** | 0.09977 |
| XV | $a_0/a_1$ | 0.09029 | 0.47354 | **0.13691** | 0.13452 | 0.14334 |
| XVI | A | **0.31396** | **0.83644** | 0.085518 | 0.083247 | **0.24818** |

Table 2 – Results for transformations to generalized axis for problems with linearly nonseparable classes

| $\aleph_z(x_i^s, x_j^s)$ | Designations at figures | Averaged estimations of criterions | | | | |
|---|---|---|---|---|---|---|
| | | $I_z$ | $I_z'$ | $r_z^{зн.}$ | $r_z^{Ф.}$ | $r_z$ |
| $x_1$ | $x_1$ | 0.28329 | 0.31701 | **0.70166** | **0.69245** | 0.70465 |
| $x_2$ | $x_2$ | 0.059362 | 0.1273 | 0.31313 | 0.30484 | 0.33693 |
| I | $x_1+x_2$ | 0.17709 | 0.39723 | 0.5955 | 0.58127 | 0.65885 |
| II | $x_1x_2$ | 0.080369 | 0.1707 | 0.53548 | 0.48836 | 0.55709 |
| III | $x_1^2+x_2^2$ | 0.099449 | 0.21696 | 0.5709 | 0.56416 | 0.62864 |
| IV | projection | **0.66056** | **0.72692** | **0.85241** | **0.83999** | **0.82336** |
| V | d | 0.4924 | **0.66342** | 0.29445 | 0.29213 | 0.24224 |
| VI | a | 0.049716 | 0.10082 | 0.17121 | 0.15757 | 0.19804 |
| VII | a·d | 0.050559 | 0.096714 | 0.094624 | 0.11141 | 0.077042 |
| VIII | a/d | 0.076551 | 0.1392 | 0.27185 | 0.26625 | 0.28498 |
| IX | d·cos(a) | 0.24959 | 0.29058 | 0.64768 | 0.63231 | 0.5295 |
| X | $d_0$ | 0.11626 | 0.19528 | 0.2476 | 0.253 | 0.25659 |
| XI | $a_0$ | 0.17957 | 0.37984 | 0.70164 | 0.68825 | **0.75375** |
| XII | $d_1$ | 0.16503 | 0.36229 | 0.26152 | 0.24004 | 0.20995 |
| XIII | $a_1$ | 0.087524 | 0.14402 | 0.5367 | 0.53116 | 0.60528 |
| XIV | $d_0/d_1$ | 0.069269 | 0.12202 | 0.23033 | 0.21522 | 0.2116 |
| XV | $a_0/a_1$ | 0.66165 | **0.77766** | **0.82875** | **0.81252** | **0.84694** |
| XVI | A | 0.44263 | 0.5511 | 0.26628 | 0.26681 | 0.25541 |

# LOCALITY SENSITIVE HASHING

The hash for *s*-th instance is determined by the formula:

$$x_*^s = \sum_{j=1}^{N} w_j x_j^s ,$$

where

$x_j^s$ is a value of the *j*-th input feature of *s*-th instance in a sample;

$w_j$ is a weight of *j*-th feature;

*N* is a number of features characterizing sample.

The weight values are obtained by the specific method.

Usually, the weights are obtained by iterative optimization as specific selection of random transformation with best properties.

# TABLE HASHING

## Hash computing

1. Normalize feature values and map to the integer numbers in the range from 0 to $2^L$ (where $L$ is a computer bit grid size):

$$x_j^{s*} = \left\lfloor \left( \frac{x_j^s - \min\limits_{i=1,2,..,N} \{x_i^s\}}{\max\limits_{i=1,2,..,N} \{x_i^s\} - \min\limits_{i=1,2,..,N} \{x_i^s\}} \right) 2^L \right\rfloor .$$

2. Compress feature values by transforming from $L$ digit numbers to $q$ digit numbers:

$$x_j^{s*} = x_j^{s*} \operatorname{div} 2^{L-q},$$

where $1 \le q \le \lceil \log_2 K \rceil < L$.

3. Compute hash for each $s$-th instance:

$$x_*^s = \sum_{j=1}^{N} w_j x_j^{s*}, \; s = 1, 2, ..., S.$$

## Weights computing

For a hash taking into account only first $L / q$ features:

$$w_j = \begin{cases} 2^{L-(q+1)(j-1)}, j \le L \operatorname{div} q; \\ 0, j \ge L \operatorname{div} q. \end{cases}$$

For a hash taking into account all features:

$$w_j = \begin{cases} 2^{L-(q+1)(j-1)}, j \le L \operatorname{div} q; \\ 2^{L-(q+1)(p-1)}, p = j \bmod (L \operatorname{div} q), j > L \operatorname{div} q; \end{cases}$$

## LOCALITY SENSITIVE FEATURE TRANSFORMATION WITH A FEATURE RANKING

To exclude the sorting of random projection of a sample from original feature space we will **consider hierarchy of feature space partitions** on regions, replacing real feature values on discrete (integer) numbers of intervals on the feature axis **aiming** for each feature to find such partitioning on intervals at which **the number of intervals will be minimal providing required accuracy**.

**Stage of Initialization.** Set the original sample <*x*, *y*>, and maximum allowable error value $0 \le \varepsilon << S$. Normalize all feature values:

$$x_j^s = \frac{x_j^s - \min\limits_{i=1,2,..,N}\{x_i^s\}}{\max\limits_{i=1,2,..,N}\{x_i^s\} - \min\limits_{i=1,2,..,N}\{x_i^s\}}.$$

**Stage of feature partitioning limitations definition.** Define the maximal number of equal size intervals *Q*, on which the feature ranges can be partitioned:

$$Q = S \text{ or } Q = \max\{\lceil \log_2 S \rceil, \min\{S, 2K\}\},$$

where *K* is a number of classes, *S* is a number of instances in a sample.

For each $j$ =1, 2, ..., $N$ set the number of intervals, on which the $j$-th feature range will be partitioned: $Q_j = Q$.

**Stage of feature partitioning.** For each *j*-th feature, *j* =1, 2, ..., *N* sequentially execute steps 1–4.

1. Split the range of *j*-th feature on $Q_j$ intervals of equal size.

2. For each *q*-th interval of *j*-th feature, *q*=1,2,..., $Q_j$, determine:

– the number of instances of *k*-th class, hitting to it $S_k^{j,q}$, *k* = 1, 2, ..., *K*;

– the error for each *q*-th interval of *j*-th feature:

$$E_{j,q} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{1 \mid -1 \le Qx_j^s - q \le 0, -1 \le Qx_j^p - q \le 0, y^s \ne y^p\}.$$

3. Evaluate summary error for all intervals of *j*-th feature: $E_j = \sum_{q=1}^{Q_j} E_{j,q}$.

4. If $E_j \le \varepsilon$ then set: $Q_j = \lceil Q_j / 2 \rceil$ and go to step 1, otherwise return the previous partition of the *j*-th feature range.

**Stage of feature ranking.** For *j* =1, 2, ..., *N* define feature ranks $r_j$ in the order of $Q_j$ increasing: the bigger the $Q_j$ the less the rank of *j*-th feature. For the features with equal values of $Q_j$ we heuristically assume that the most important (with a bigger rank) is the feature, which individually is more significant for the output variable or the feature having less number.

**Stage of weights computation.** Set the weights:

$$w_j = \left( \max_{j=1,2,...,N} \{Q_j\} \right)^{N-r_j} \quad \text{or} \quad w_j = \left( 2^{\left\lceil \log_2 \max_{j=1,2,...,N} \{Q_j\} \right\rceil} \right)^{N-r_j}$$

# THE HEURISTIC TRANSFORMATIONS TO THE GENERALIZED AXIS

For large-scale problems it is advisable to ensure the creation of such transformations, which would allow the mapping of individual instances without loading of whole initial sample, as well as taking into account the feature informativity in the process of transforming and to provide a generalization of data.

To **ensure the generalization of close located data points** (instances) we propose to replace feature values to numbers of feature value interval. For this we need previously to discretize the features by partitioning them into intervals of values.

**To partitioning the features into intervals** the number of interval (term), which hits the *s*-th instance on the *j*-th feature is proposed to determine as:

$$\hat{x}_j^s = \begin{cases} \text{round}\left(1 + \dfrac{x_j^s - x_j^{\min}}{\theta_j}\right), \theta_j > 0; \\ 1, \theta_j = 0, \end{cases}$$

$$\theta_j = \frac{x_j^{\max} - x_j^{\min}}{k_j},$$

$$k_j = \begin{cases} K, K > \text{round}(\ln S), K < \sqrt[N]{S}; \\ \max\{2, \text{round}(\sqrt[N]{S})\}, K > \text{round}(\ln S), K > \sqrt[N]{S}; \\ \max\{2, \text{round}(\ln S)\}, K < \text{round}(\ln S) < \sqrt[N]{S}; \\ K, \text{round}(\ln S) \le K, K < \sqrt[N]{S}; \\ \max\{2, \text{round}(\sqrt[N]{S})\}, \text{round}(\ln S) \le K, K \ge \sqrt[N]{S}, \end{cases}$$

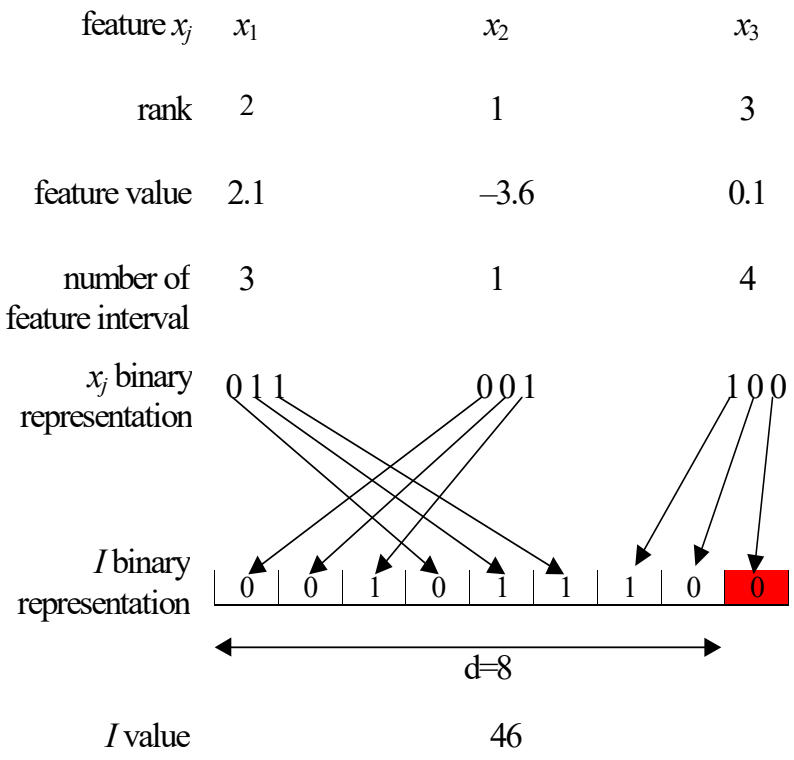
where $x_j^{\min}$, $x_j^{\max}$ are the minimum and maximum values of *j*-th feature, respectively.

## Transformation 1.

For each number of interval of $j$-th feature get it binary representation (binary numbers padded with zeros from the left to $c_j$ – the number of digits in $k_j$).

Set the coordinate of $s$-th instance on the generalized axis $I^s$=0, set the position (bit) number of generalized axis coordinate $p$ = 1.

Going by the feature numbers $j$ in descending order of their rank and by the group of digits in the interval number $c$ = 1, 2 , ..., $c_j$ perform in a cycle: if $p \le d$, where $d$ is a number of bits in a computer bit grid, then record at $p$-th position (with numbering from left) of the binary representation of a generalized feature $I^s$ the $c$-th position value (with numbering from left) of interval number, in which the $s$-th instance hit on the $j$-th feature, and set: $p$=$p$+1.




As a result, we will obtain a generalized axis coordinate of instance with the implicit ranking and selection of features.

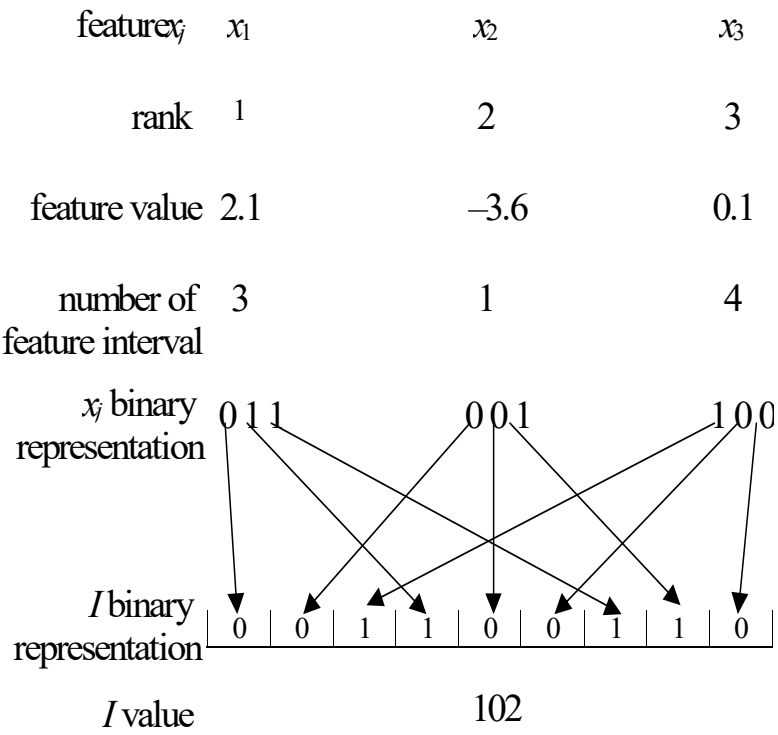
## Transformation 2.

It is an alternative format of constructing a generalized feature for transformation 1.

If the total number of bits to represent interval numbers of all features $c_j k_j N$ does not exceed the number of bits in a bit grid $d$ when the values $c_j$ are equal for all features:

– for each interval number of $j$-th feature obtain its binary representation (binary numbers padded with zeros from the left to $c_j$ – the number of digits in $k_j$), set the coordinate of $s$-th instance on the generalized axis $I^s = 0$, set the position number of coordinate on a generalized axis $p = 1$;

– looking in a cycle on a group of digits in the interval number $c = 1, 2, ..., c_j$ and on feature numbers $j$ in the descending order of their ranks: put to the $p$-th bit position (numbering from the left) of the binary representation of the generalized feature $I^s$ the $c$-th bit (numbering from the left) of interval number, in which the $s$-th instance hits on the $j$-th feature and set: $p=p+1$.




As a result, we will obtain the generalized axis coordinate with implicit ranking of features.

**Transformation 3.**

The generalized feature formed on the basis of ***locality-preserving hashing***.

The initial feature space is divided into $2^k$ equal hypercubes, each of which identified by the key $I^s$ of a $k$ bit length, where $k$ is a number of feature partitions.

After the $i$-th partition the initial feature space split to $2^i$ $N$-dimensional cubes, wherein the $i$-th partition is carried out on the $j$-th dimension: $j = i \bmod N$.

At the $i$-th partition if hypercube located in the top half of the partitioned range, then set to one the $i$-th bit of its key, and otherwise set the $i$-th bit of its key to zero (set to one the bit in the $i$-th position of $k$-bit identifier, extended by zeros from the left, if the length is less than $k$).

The key $I^s$ algorithmically can be generated as follows:

set: $I^s = 0$, $x_j^{\min'} = x_j^{\min}$, $x_j^{\max'} = x_j^{\max}$,

then for $i = 1, 2, ..., k$ do:

set: $j = i \bmod N$, $x_j^{\text{mid}} = (x_j^{\min'} + x_j^{\max'})/2$, $I^s = 2I^s$;

if $x^s_j > x_j^{\text{mid}}$ then set: $x_j^{\min'} = x_j^{\text{mid}}$, $I^s = I^s + 1$,

else set: $x_j^{\max'} = x_j^{\text{mid}}$.

**Transformation 4.**

The above-described transformations provide mapping to the discrete generalized axis.

If the total number of bits to represent numbers of all feature intervals exceeds the number of bits in a bit grid of computer, it is possible to use a transformation to the generalized real axis with partial information loose:

add to the real coordinate on generalized feature $I^s$ the $c$-th bit (numbering from the left) of interval number, in which the $s$-th instance hits on the $j$-th feature:

$$I^s = \sum_{j=1}^{N} \frac{w_j \hat{x}_j^s}{r_j k_j},$$

$$w_j = \frac{1}{k_j} \sum_{k=1}^{k_j} w_{j,k}\ ,\ w_{j,k} = \left\{ \frac{\max\limits_{q=1,2,..,K} \{S_{j,k}^q\}}{S_{j,k}} \middle| S_{j,k} > 0 \right\},\ S_{j,k} = \sum_{q=1}^{K} S_{j,k}^q,$$

where $S_{j,k}^q$ is a number of instances of $q$-th class located in the $k$-th interval of $j$-th feature,

$r_j$ is a rank of $j$-th feature (the number of $j$-th feature in decreasing order of individual feature importance).

**Transformation 5.**

Define the distance from the *s*-th instance to the unit vector in the normalized coordinate system:

$$d^s = \sqrt{\sum_{j=1}^{N} \left(\hat{x}_j^s - 1\right)^2},$$

and the angle between the instance as a vector and the unit vector:

$$\varphi^s = \arccos\left( \frac{\sum_{j=1}^{N} \hat{x}_j^s}{\sqrt{N \sum_{j=1}^{N} \left(\hat{x}_j^s\right)^2}} \right).$$

Thus we map the *s*-th instance from the *N*-dimensional space into two-dimensional space.

Next for coordinates of *s*-th instance in formed two-dimensional space by analogy with the first transformation obtain coordinate of *s*-th instance on the generalized axis $I^s$.

**Transformation 6.**

Generate $Q$ support vectors – the centers of pseudo-clusters $C^q = \{C^q_j\}$, $q$=1, 2, ..., $Q$, $K \le Q << S, j = 1, 2, ..., N$.

In the simplest case their coordinates can be set as random taking into account dimensionality and feature scales ($x^{\min}{}_j \le C^q_j \le x^{\max}{}_j$), or by setting $Q$=$K$ to determine the center of each its class:

$$C^q_j = \frac{1}{S^q} \sum_{s=1}^{S} \{x^s_j \mid y^s = q\},$$

$$j = 1, 2, ..., N, q=1, 2, ..., K.$$

After this calculate the clusters based on their proximity and position in feature space relative to the smallest feature values:

– find the distance from the cluster centers to the point with the lowest feature values:

$$R_{\min}(C^q) = \sum_{j=1}^{N} \left(C^q_j - x^{\min}_j\right)^2.$$

– find the distance between the cluster centers:

$$R(C^q, C^p) = \sum_{j=1}^{N} \left(C^q_j - C^p_j\right)^2.$$

– find the center of cluster closest to the point with the lowest feature values:

$$q = \arg \min_{g=1,2,...,Q} \{R_{\min}(C^g)\};$$

– set this center as the current, set a new number of current cluster $t$=1, put current cluster in the set of centers with a new index ($C^*$=$C^* \cup C^{*1}$, $C^{*1} = C^q$) and delete it from the set of centers without a new index ($C$=$C / C^q$);

– while exist at least one cluster without a new index (i.e. $C \ne \varnothing$) perform: among the remaining clusters without a new index in $C$ find the closest cluster to the current cluster:

$$p = \arg \min_{\substack{g=1,2,...,Q; \\ C^g \in C}} \{R(C^q, C^g)\}.$$

then increase $t = t$+1, put the current cluster to the set of centers with a new index ($C^*$=$C^* \cup C^{*t}$, $C^{*t} = C^p$) and remove it from the set of centers without a new index ($C$=$C / C^p$).

As a result we will receive $C^*$ – a set of cluster centers with numbers corresponding to their proximity to the point with the lowest values of features, and also allowing to determine qualitatively the proximity of the cluster centers.

Further for each instance of the initial sample $x^s$, $s$=1, ..., $S$ do:

– define the distance from it to each cluster center, $q$=1,2,...,$Q$, by:

$$R(x^s, C^{*q}) = \sum_{j=1}^{N} \left(x_j^s - C_j^{*q}\right)^2;$$

– find the index of the nearest cluster center:

$$p = \arg \min_{q=1,2,...,Q} \{R(x^s, C^{q*})\};$$

– find the angle between the vectors $x^s$ and $C^{p*}$ relative to the point with the lowest feature values:

$$\varphi = \arccos \frac{\sum_{j=1}^{N} (x_j^s - x_j^{\min})(C_j^{p*} - x_j^{\min})}{\sqrt{\sum_{j=1}^{N} \left(x_j^s - x_j^{\min}\right)^2} \sqrt{\sum_{j=1}^{N} \left(C_j^{p*} - x_j^{\min}\right)^2}};$$

– assign the $s$-th instance with the coordinate on the generalized axis:

$$I^s = p + \frac{\varphi}{\pi}.$$

## THE CHARACTERISTICS OF TRANSFORMATIONS TO THE GENERALIZED AXIS

**The characteristics of instance mapping process:**

– $t^s$ is a time of transforming of one instance from the original feature space to the generalized axis for the sequential computations;

– $m^s$ is a computer memory volume (size) used by the transformation method for processing one instance;

– $\lambda$ is a number of adjustable parameters of transformation needed for its implementation;

– $t$ is a time of calculation of transformation parameters based on the training sample;

– $m$ is a computer memory volume used to calculate the transformation parameters on the basis of training sample.

# Collisions

| | |
|---|---|
| **Definition.** ***Collision*** is a situation where several instances have equal coordinates may occur in the original and in the synthesized feature spaces.<br><br>**Definition.** ***Collision point*** is a point in the feature space, in which there is a collision. | **Effect:** The collision is quite admissible and even desirable in problems of automatic classification on condition that all instances located at the point of collision belongs to the same class. However, if the instances located at the point of collision, belongs to different classes, the used feature set does not provide a good separability of instances. |

**Notation:**

$\{g_v\}$ is a set of points of collision, $v$=1, 2, ..., $V$,

$g_v$ is a set of instances belonging to a $v$-th point of collision,

$V$ is the number of points of collision, which obviously can not exceed 0,5$S$.

## The indicators to estimate the quality of the results of transformations

**The number of points of collisions** in which instances belongs to different classes, after the transformation of the sample to the generalized axis:

$$E^*_{<I,y>} = \sum_{v=1}^{V} \{1 \mid \exists s, p = 1,2,...,S, p \neq s : I^s \in g_v, I^p \in g_v, I^s = I^p, y^s \neq y^p\}$$

This indicator in the best case will be equal to zero when there is no collision points, and in the worst case it maximum value will not exceed 0,5*S*.

**The probability estimation (frequency) of the collision points** in which instances belonging to different classes, after the transformation of the sample to the generalized axis:

$$P^*_{<I,y>} = \frac{2E^*_{<I,y>}}{S}.$$

**The corrected number of points of collisions** in which instances belong to different classes, after the transformation of the sample to the generalized axis:

$$E^{*'}_{<I,y>} = E^*_{<I,y>} - E^*_{<x,y>},\ E^*_{<x,y>} = \sum_{v=1}^{V} \left\{ 1 \left| \begin{array}{c} \exists s, p = 1,2,...,S, p \neq s : x^s \in g_v, x^p \in g_v, \\ y^s \neq y^p, \forall j = 1,2,...,N : x^s_j = x^p_j \end{array} \right. \right\}$$

where $E^*_{<x,y>}$ – is the number of collision points in which the instances belongs to different classes in the initial sample. The indicator $E^{*'}_{<I,y>}$ more accurately characterizes the quality of transformation to the generalized axis because it eliminates the errors present in the sample. In the best case it will be equal to zero when there is no collisions, and in the worst case it maximum value will not exceed 0,5*S*.

**The corrected probability estimation (frequency) of the collision points** in which the instances belong to different classes after the sample transformation to the generalized axis:

$$P^{*'}_{<I,y>} = \frac{2E^{*'}_{<I,y>}}{S}.$$

**The total number of instances in the collision points** in which the instances belong to different classes after the sample transformation to the generalized axis:

$$E^{\Sigma}_{<I,y>} = \sum_{v=1}^{V} \left\{ |g_v| \left| \begin{array}{l} \exists s, p = 1,2,...,S, p \neq s : I^s \in g_v, \\ I^p \in g_v, I^s = I^p, y^s \neq y^p \end{array} \right. \right\}$$

The more will be value of this indicator, the worse separability of instances on the generalized axis. In the best case it will be equal to zero and in the worst case it will not exceed the number of instances in the sample *S*.

**The probability estimation of instance hitting to the collision point** in which instances belong to different classes after the sample transformation to the generalized axis:

$$P^{\Sigma}_{<I,y>} = \frac{E^{\Sigma}_{<I,y>}}{S}.$$

**The total number of instances in the collision points** of the initial sample in which the instances belong to different classes:

$$E^{\Sigma}_{<x,y>} = \sum_{v=1}^{V} \left\{ |g_v| \left| \begin{array}{l} \exists s, p = 1,2,...,S, p \neq s : x^s \in g_v, x^p \in g_v, \\ y^s \neq y^p, \forall j = 1,2,...,N : x^s_j = x^p_j \end{array} \right. \right\}.$$

The more will be the value of this indicator, the worse the separability of instances of the initial sample will be. In the best case it will be equal to zero and in the worst case it not exceed the number of instances in the sample *S*.

**The probability estimation of instance collision in the sample** in which instances belong to different classes:

$$P^{\Sigma}_{<x,y>} = \frac{E^{\Sigma}_{<x,y>}}{S}.$$

**The number of pairwise collision of instances of different classes** after the sample transformation to the generalized axis:

$$E_{<I,y>} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{ y^s \neq y^p \mid I^s = I^p \}.$$

In the best case, this indicator is zero when there is no any collision, and in the worst case its value will not exceed $S(S-1)$.

**The probability estimation (frequency) of pairwise collision of instances of different classes** after the sample transformation to the generalized axis:

$$P_{<I,y>} = \frac{E_{<I,y>}}{S(S-1)}.$$

**The corrected number of pairwise collision of instances of different classes** after the transformation of training and (or) test sample to the generalized axis:

$$E'_{<I,y>} = E_{<I,y>} - E_{<x,y>}, \quad E_{<x,y>} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{y^s \neq y^p \mid \forall j = 1,2,...,N : x_j^s = x_j^p\},$$

where $E_{<x,y>}$ is a number of pairwise collision of instances of different classes in the original sample.

This indicator $E'_{<I,y>}$ in comparison with the previous indicator more accurately characterizes the quality of the transformation to the generalized axis, because it eliminates the errors present in the original sample. In the best case, it would be equal to zero when there is no any collision, and at worst case, it maximum value will not exceed $S(S–1)$.

**The corrected probability estimation (frequency) of pairwise collisions of instances of different classes** after sample transformation to the generalized axis:

$$P'_{<I,y>} = \frac{E'_{<I,y>}}{S(S-1)}.$$

**The average number of clusters per class on a generalized axis:**

$$\bar{k} = \frac{k}{K}.$$

where $k$ is a number of clusters of different classes on a generalized axis.

To determine $k$, we need order the instances $<I^s, y^s>$ in ascending order on the generalized axis. Then, looking from left to right, we need to identify clusters – the intervals of one-dimensional axis, all instances of each of which belong to only one class.

The less will be the number of such clusters, the simply is partition of generalized axis. In the best case when the classes are compact, i.e. $k = K$, this indicator is equal to one. The more will be value of this indicator, the worse the separability of instances will be on the generalized axis. In the worst case where each instance falls into a single cluster its value will be $\bar{k} = S / K$.

| | |
|---|---|
| **The minimum distance** between instances of different classes on the generalized axis:<br><br>$$R'_{\min} = \min_{\substack{s=1,\ldots,S;\\ p=s+1,\ldots,S}} \left\{ \| I^s - I^p \| \, \middle| \, y^s \neq y^p \right\}.$$<br><br>The more will be value of this ratio, the better classes will be separated on the generalized axis. | **The maximum distance** between instances of one class on the generalized axis:<br><br>$$R'_{\max} = \max_{\substack{s=1,\ldots,S;\\ p=s+1,\ldots,S}} \left\{ \| I^s - I^p \| \, \middle| \, y^s = y^p \right\}.$$<br><br>The less will be this indicator value, the more compact instances of each class will be positioned on the generalized axes. |

**The average ratio of distances** on the generalized axis and in the original feature space:

$$\overline{\Delta} = \frac{R_{\max}}{0{,}5S(S-1)R^*_{\max}} \sum_{s=1}^{S} \sum_{p=s+1}^{S} \left\{ \frac{|I^s - I^p|}{R(x^s, x^p)} \middle| R(x^s, x^p) > 0 \right\},$$

where $R^*_{\max} = \max\limits_{\substack{s=1,\ldots,S;\\ p=s+1,\ldots,S}} \left\{ |I^s - I^p| \right\}$, $R_{\max} = \max\limits_{\substack{s=1,\ldots,S;\\ p=s+1,\ldots,S}} \left\{ R(x^s, x^p) \right\}$, $R(x^s, x^p) = \sqrt{\sum_{j=1}^{N} (x_j^s - x_j^p)^2}$.

The more will be value of this indicator, the better on the average the transformation on the generalized axis reflects location of instances in the original space and features the better separability of instances on the generalized axis.

**The average of the relative distance products** on the generalized axis and in the original feature space:

$$\Delta = \frac{\sum_{s=1}^{S} \sum_{p=s+1}^{S} |I^s - I^p| R(x^s, x^p)}{0{,}5S(S-1)R^*_{\max} R_{\max}}.$$

This indicator will vary from zero to one: The more will be its value, the better on the average the transformation on the generalized axis reflects the location of instances in the original feature space.

**The indicator of generalized axis feasibility of establishing:**

$$G = \frac{\min\limits_{j=1,2,\ldots,N} \left\{ k_{<x_j, y>} \right\}}{k},$$

where $k_{<x_j, y>}$ is the number of intervals of different classes on the axis of feature $x_j$.

This indicator in the best case will be equal to $S / K$, and in the worst case will be equal to $K / S$. If this indicator will be greater than one, the use of the generalized axis will be feasible, otherwise it can be replaced with the original feature, characterized by the smallest number of intervals of different classes.

# THE COMPARISON CRITERIA OF GENERALIZED AXIS TRANSFORMATIONS

**The combined criterion of the minimum of time and memory on the instance transformation:**

$$F_1 = t^s m^s \to \min.$$

**The combined criterion of the minimum of time and memory to determine the transformation parameters for the training sample:**

$$F_2 = \lambda\, tm \to \min.$$

**The integral criterion:**

$$F_3 = t^s m^s + \lambda S^{-1} tm \to \min.$$

**The criterion of the minimum of probability of instance group collisions:**

$$F_4 = P'_{<I,y>} \to \min.$$

**The criterion of the minimum of probability of instance pair collisions:**

$$F_5 = P^{*'}_{<I,y>} \to \min.$$

**The combined criterion of the minimum probability of pair and group collisions:**

$$F_6 = 0.5(P'_{<I,y>} + P^{*'}_{<I,y>}) \to \min.$$

**The maximum of class compactness-separability:**

$$F_7 = \bar{k} \to \min.$$

**The integral criterion of minimum of collisions-compactness-separability of classes:**

$$F_8 = \frac{\bar{k}}{2}\left(P'_{<I,y>} + P^{*'}_{<I,y>}\right) \to \min,\ \bar{k} > 0.$$

**The integral criterion of minimum of collision-maximum of compactness-separability of classes and maximum of average of relative distance products on the generalized axis and in the original feature space:**

$$F_9 = \frac{\bar{k}\left(P'_{<I,y>} + P^{*'}_{<I,y>}\right)}{1 + \Delta e^{-\bar{\Delta}+1}} \to \min.$$

**The integral criterion of the minimum of collisions-maximum of a generalized axis establishing feasibility-compactness-separability of classes and the maximum of average of relative distances products on the generalized axis and in the original feature space:**

$$F_{10} = \frac{\left(P'_{<I,y>} + P^{*'}_{<I,y>}\right)}{G + \Delta e^{-\bar{\Delta}+1}} \to \min.$$

## THE EXPERIMENTS AND RESULTS

### The characteristics of initial data samples and the fragment of the experimental results to study the transformations on generalized axis

| Characteristics | Task | | | |
|---|---|---|---|---|
| | ***Gas-turbine air-engine blade diagnosis*** | ***Chronic obstructive bronchitis diagnosis*** | ***Agricultural plant recognition on the remote sensing data*** | ***Fisher Iris classification*** |
| $S$ | 32 | 205 | 3226 | 150 |
| $N$ | 513 | 28 | 256 | 4 |
| $K$ | 2 | 2 | 3 | 3 |
| ***Best transformation:*** | | | | |
| number | 2 | 1 | 4 | 1 |
| $E^*_{<I,y>}$ | 1 | 0 | 0 | 1 |
| $P^*_{<I,y>}$ | 0.0625 | 0 | 0 | 0.013333 |
| $E_{<I,y>}$ | 1 | 0 | 0 | 9 |
| $P_{<I,y>}$ | 0.0010081 | 0 | 0 | 0.00040268 |
| $E^*_{<x,y>}$ | 0 | 0 | 0 | 0 |
| $E^{*'}_{<I,y>}$ | 1 | 0 | 0 | 1 |
| $E_{<x,y>}$ | 0 | 0 | 0 | 0 |

| Best transformation: | Task | | | |
|---|---|---|---|---|
| | Gas-turbine air-engine blade diagnosis | Chronic obstructive bronchitis diagnosis | Agricultural plant recognition on the remote sensing data | Fisher Iris classification |
| $E'_{<I,y>}$ | 1 | 0 | 0 | 9 |
| $E^{\Sigma}_{<x,y>}$ | 0 | 0 | 0 | 0 |
| $P^{\Sigma}_{<x,y>}$ | 0 | 0 | 0 | 0 |
| $E^{\Sigma}_{<I,y>}$ | 2 | 0 | 0 | 10 |
| $P^{\Sigma}_{<I,y>}$ | 0.0625 | 0 | 0 | 0.066667 |
| $k$ | 15 | 72 | 1773 | 11 |
| $R_{\max}$ | 70.114 | 918.99 | 7.5551 | 7.0852 |
| $R^{*}_{\max}$ | $1.5228 \cdot 10^{9}$ | $1.6927 \cdot 10^{9}$ | 3.3217 | $1.5126 \cdot 10^{9}$ |
| $R'_{\min}$ | 0 | 1056 | $1.1437 \cdot 10^{-7}$ | 0 |
| $R'_{\max}$ | $1.5228 \cdot 10^{9}$ | $1.6833 \cdot 10^{9}$ | 3.318 | 609746944 |
| $\overline{\Delta}$ | 0.82534 | 0.69432 | 0.37993 | 0.88456 |
| $\Delta$ | 0.095099 | 0.098493 | 0.028541 | 0.18511 |
| $G$ | 0.2 | 0.69444 | 0.56007 | 1 |
| $t^{s}$ | 0.00097501 | 0.00053269 | $3.8686 \cdot 10^{-5}$ | 0.000416 |
| $m^{s}$ | 4309 | 262.83 | 2057.9 | 82.133 |

| Best transformation: | Task | | | |
|---|---|---|---|---|
| | Gas-turbine air-engine blade diagnosis | Chronic obstructive bronchitis diagnosis | Agricultural plant recognition on the remote sensing data | Fisher Iris classification |
| $t$ | 0.2184 | 0.093601 | 4.524 | 0.1092 |
| $m$ | 333484 | 109644 | 6764132 | 20340 |
| $\lambda$ | 1539 | 84 | 1024 | 12 |
| $F_1$ | 4.2013 | 0.14001 | 0.079613 | 0.034168 |
| $F_2$ | $1.1209 \cdot 10^8$ | $8.6207 \cdot 10^5$ | $3.1336 \cdot 10^{10}$ | $0.26654 \cdot 10^5$ |
| $F_3$ | $3.5028 \cdot 10^6$ | 4205.4 | $9.7134 \cdot 10^6$ | 177.73 |
| $F_4$ | 0.0010081 | 0 | 0 | 0.00040268 |
| $F_5$ | 0.0625 | 0 | 0 | 0.013333 |
| $F_6$ | 0.031754 | 0 | 0 | 0.006868 |
| $F_7$ | 7.5 | 36 | 591 | 3.6667 |
| $F_8$ | 0.23816 | 0 | 0 | 0.025183 |
| $F_9$ | 0.42786 | 0 | 0 | 0.041701 |
| $F_{10}$ | 0.20274 | 0 | 0 | 0.011373 |

# POLAR COORDINATES BASED HASHING

**Stage of initialization.**

Set the original sample <*x*, *y*>.
Normalize feature values:

$$x_j^s = \frac{x_j^s - \min\limits_{i=1,2,..,N}\{x_i^s\}}{\max\limits_{i=1,2,..,N}\{x_i^s\} - \min\limits_{i=1,2,..,N}\{x_i^s\}}.$$

**Stage of coordinates transformation to the polar system.**

Define polar coordinates of instances $<\{<\rho^s, \varphi^s>\}, y>$, where $\varphi^s = \{\varphi_j^s\}$.
Radial coordinate for *s*-th instance:

$$\rho^s = \sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2}.$$

Angles of *s*-th instance relative to coordinate axes in the original feature coordinate system in radians:

$$\varphi_j^s = \arccos\frac{x_j^s}{\sqrt{\sum\limits_{i=j}^{N}\left(x_i^s\right)^2}}, j = 1, 2, ..., N–1.$$

Then reduce data description by converting angle coordinates to integer values:

$$\varphi_j^s = \left\lceil \frac{\varphi_j^s 180}{\pi} \right\rceil.$$

**Stage of hash transformation forming.**

Using one of the further described methods compute the hashes for all instances in a polar system.

**Method 1.**

Hash computed as

$$x_*^s = w^{\rho}\rho^s + \sum_{j=1}^{N-1} w_j^{\varphi}\varphi_j^s ,$$

$$w^{\rho} = 2^{L-P+1},\ w_j^{\varphi} = 2^{L-\mathrm{P}-\Phi j+1},$$

$$\rho^s = \left\lfloor \frac{\rho^s}{\frac{\sqrt{N}}{2^{\mathrm{P}}}} \right\rfloor = \left\lfloor \frac{\sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2}}{\frac{\sqrt{N}}{2^{\mathrm{P}}}} \right\rfloor = \left\lfloor \frac{2^{\mathrm{P}}}{\sqrt{N}}\sqrt{\sum_{j=1}^{N}\left(x_j^s\right)^2} \right\rfloor,$$

$$\varphi_j^s = \left\lfloor \frac{\varphi_j^s}{\frac{90^\circ}{2^{\Phi}}} \right\rfloor = \left\lfloor \frac{2^{\Phi}\varphi_j^s}{90^\circ} \right\rfloor,$$

$$\Phi = \left\lfloor \frac{L-\mathrm{P}}{N-1} \right\rfloor,\ N>1,\ \mathrm{P} = \left\lceil \log_2 \max(K,S) \right\rceil,$$

where $L$ is a computer bit grid length ($L$ = 64 for most modern computers)

**Method 2.**

Hash computed as

$$x_*^s = \sum_{g=1}^{G} 2^{(G-g)(Z+1)}\left(\rho_g^s + \sum_{j=1}^{Z} 2^{Z-j}\varphi_{jg}^s\right),$$

$$\rho_b^s = (\rho^s \bmod 2^{b+1} - \rho^s \bmod 2^b),$$

$$\varphi_{jb}^s = (\varphi_j^s \bmod 2^{b+1} - \varphi_j^s \bmod 2^b),$$

where $Z$ is a number of angles taking into account: $N-1 \ge Z \ge 1$,

$G$ is a number of bit groups: $G = \lfloor L/(Z+1) \rfloor$,

$\rho_b^s$ is a $b$-th bit value for $\rho^s$,

$\varphi_{jb}^s$ is a $b$-th bit value for $\varphi_j^s$.

**Method 3.**

Convert the polar coordinates of instances to hierarchical binary partitioning code format.

Set the number for coding *a*, its minimum $a_{\min}$ and maximum $a_{\max}$ possible values, the length of computer bit grid *L,* the initial values of code for hierarchical binary partition: $a_* = 0$, and for variable region limitations $\breve{a} = a_{\min}$, $\hat{a} = a_{\max}$.

For *i*=1, 2, ..., *L* do in a cycle: set $\bar{a} = (\breve{a} + \hat{a})/2$; if $a > \bar{a}$ then set $\breve{a} = \bar{a}$, $a_* = 2a_* + 1$, otherwise set $\hat{a} = \bar{a}$, $a_* = 2a_*$.

By analogy with a method 2, the hash of an instance is represented by a sequence of groups of bits, where in each group the first bit is the corresponding to the group of bits of the hierarchical code of an integer value or the number of the interval of the quantized distance, and the subsequent bits are the corresponding bits of codes of codes of integer values or numbers of the intervals of the angles of the instance.

**Method 4.**

Hash computed as by analogy with a method 1, but instead of the distance value or the number of its interval and instead of the values or numbers of the angle intervals, we will use their hierarchical codes, obtained similarly to method 3.

As a result, we compose the hash as follows:

the first part of the hash is a hierarchical code of the value or interval number values of the quantized distance,

the second part of the hash is a sequentially according to features the codes of integer values or numbers of intervals of values of the angles of the instance.

**Stage of hash transformation quality evaluation.**
Evaluate quality indicators for hashes formed for a sample.

The number of negative collisions:

$$N_{col-} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{1 \mid x_*^s = x_*^p, y^s \neq y_*^p\},$$

The number of positive collisions:

$$N_{col+} = \sum_{s=1}^{S} \sum_{p=s+1}^{S} \{1 \mid x_*^s = x_*^p, y^s = y_*^p\}$$

The probability of negative collisions:

$$P_{col-} = \frac{N_{col-}}{S(S-1)}$$

The probability of positive collisions:

$$P_{col+} = \frac{N_{col+}}{S(S-1)}.$$

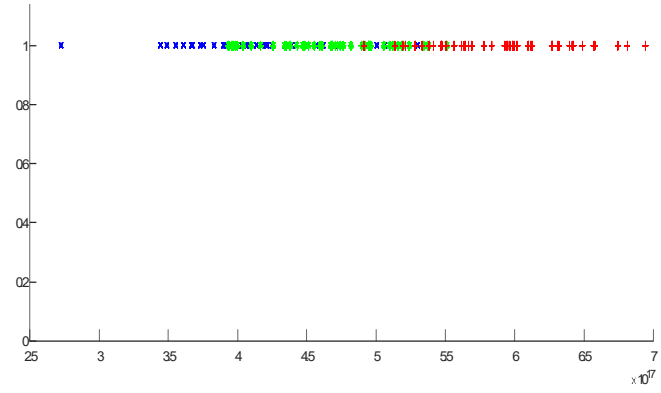
The hash will be the better the bigger probability of positive and the less probability of negative collisions.

Select as a result the best transformation.

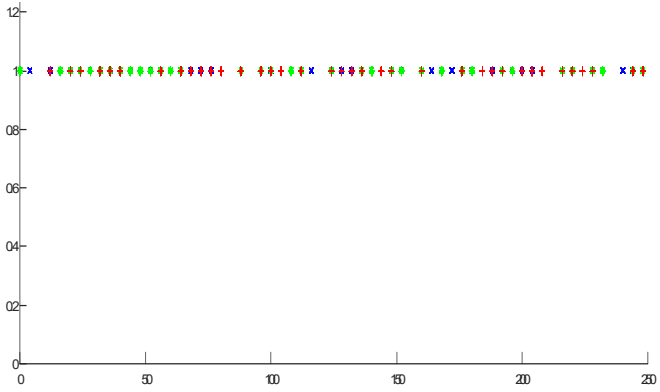
## EXPERIMENTS AND RESULTS

Table 3 – Practical problems characteristics and results of experiment

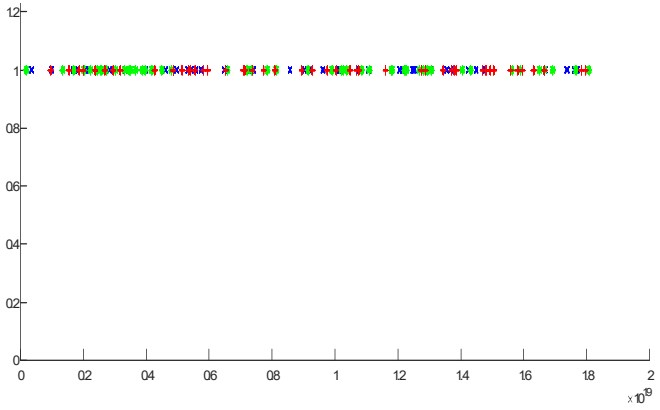
| Problem acronym | Problem name | $N$ | $S$ | $K$ | Hash computing method | $N_{col-}$ | $N_{col+}$ | $P_{col-}$ | $P_{col+}$ |
|---|---|---|---|---|---|---|---|---|---|
| Iris | Fisher Iris https://archive.ics.uci.edu/ml/datasets/iris | 4 | 150 | 2 | 1 | 0 | 1 | 0 | $4.4743\times10^{-5}$ |
| | | | | | 2 | 130 | 54 | 0.0058 | 0.0024 |
| | | | | | 3 | 0 | 1 | 0 | $4.4743\times10^{-5}$ |
| | | | | | 4 | 0 | 1 | 0 | $4.4743\times10^{-5}$ |
| Aritmia | Arrhythmia Data Set https://archive.ics.uci.edu/ml/datasets/arrhythmia | 279 | 452 | 2 | 1 | 0 | 0 | 0 | 0 |
| | | | | | 2 | 19 | 48 | $9.4036\times10^{-5}$ | $2.3756\times10^{-4}$ |
| | | | | | 3 | 8 | 20 | $3.9594\times10^{-5}$ | $9.8985\times10^{-5}$ |
| | | | | | 4 | 0 | 0 | 0 | 0 |
| Acutediag | Acute inflammation https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations | 6 | 120 | 4 | 1 | 19 | 24 | 0.0013 | 0.0017 |
| | | | | | 2 | 175 | 128 | 0.0123 | 0.0090 |
| | | | | | 3 | 19 | 24 | 0.0013 | 0.0017 |
| | | | | | 4 | 21 | 26 | 0.0015 | 0.0018 |



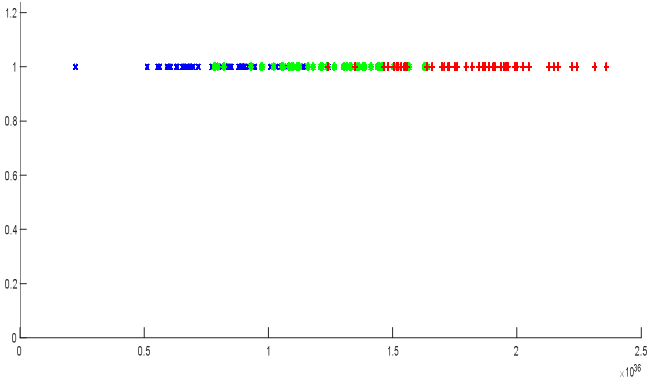
a



b

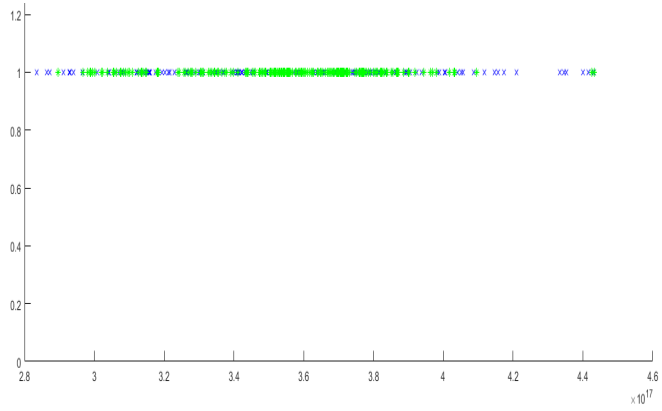


c

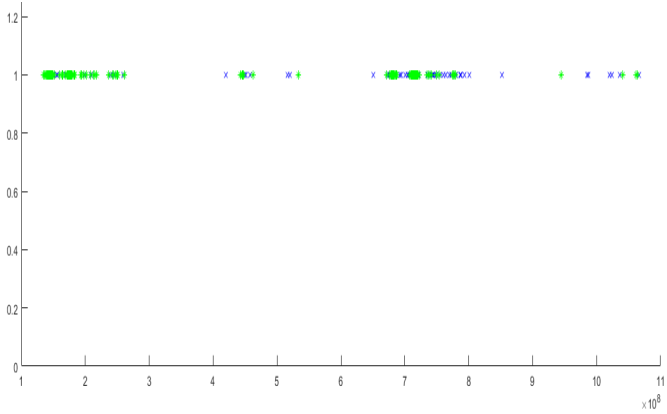


d

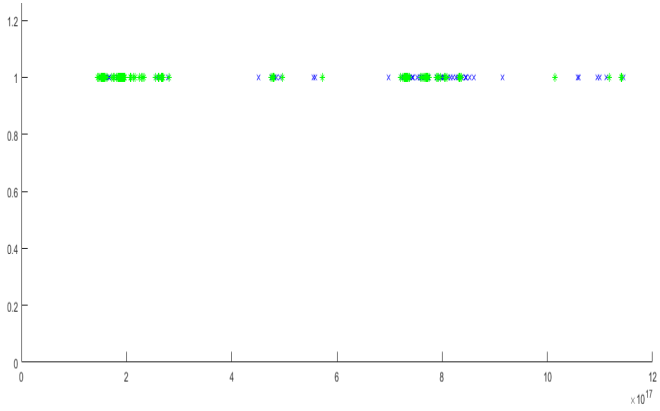
Figure 1 – Hashes for Iris problem: a –method 1, b – method 2, c – method 3, d – method 4



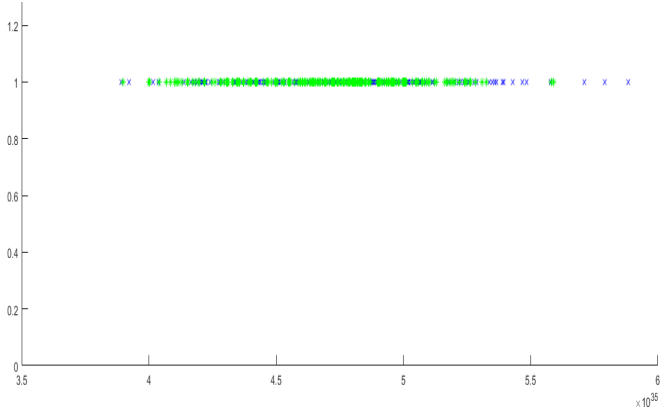

a




b

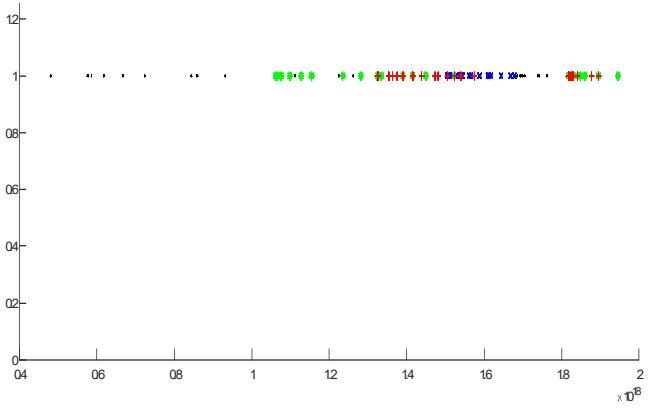



c

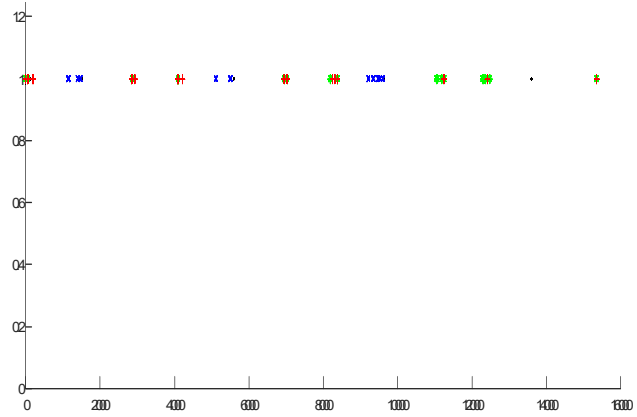



d

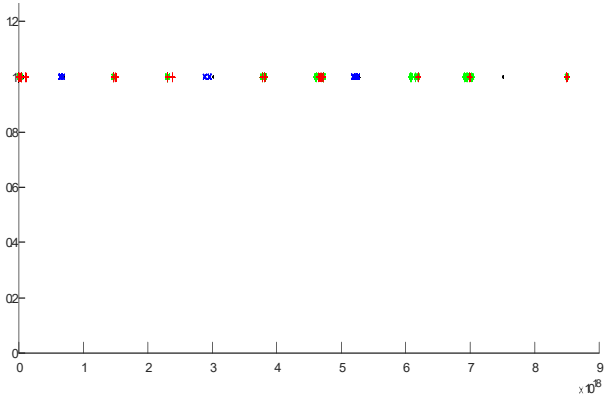
Figure 2 – Hashes for Aritmia problem: a –method 1, b – method 2, c – method 3, d – method 4



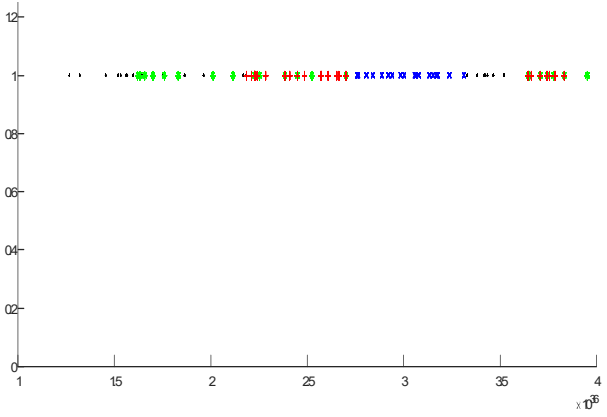
a



b



c



d

Figure 3 – Hashes for Acutediag problem: a –method 1, b – method 2, c – method 3, d – method 4

# Thank you for attention!


**Sergey A. Subbotin**

**Dr. hab. Sc., Professor,
Head of the Department
of Software Tools**


**MY CONTACTS**


**Address:**
**National University
"Zaporizhzhia Polytecnic",
Zhukovsky str., 64,
Zaporizhzhia, 69063,
Ukraine.**

**Tel.: +38-067-394-11-80,
e-mail:
subbotin@zntu.edu.ua.**