

# Error-Free Training via Information Structuring in the Classification Problem

Vladimir N. Shats\*

St. Petersburg, Russia

Email: vlshats@hotmail.com

**How to cite this paper:** Shats, V.N. (2018) Error-Free Training via Information Structuring in the Classification Problem. *Journal of Intelligent Learning Systems and Applications*, 10, 81-92.  
<https://doi.org/10.4236/jilsa.2018.103005>

**Received:** May 15, 2018

**Accepted:** August 3, 2018

**Published:** August 6, 2018

Copyright © 2018 by author and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

The present paper solves the training problem that comprises the initial phases of the classification problem using the data matrix invariant method. The method is reduced to an approximate “slicing” of the information contained in the problem, which leads to its structuring. According to this method, the values of each feature are divided into an equal number of intervals, and lists of objects falling into these intervals are constructed. Objects are identified by a set of numbers of intervals, *i.e.*, indices, for each feature. Assuming that the feature values within any interval are approximately the same, we calculate frequency features for objects of different classes that are equal to the frequencies of the corresponding indices. These features allow us to determine the frequency of any object class as the sum of the frequencies of the indices. For any number of intervals, the maximum frequency corresponds to a class object. If the features do not contain repeated values, the error rate of training tends to zero for an infinite number of intervals. If this condition is not fulfilled, a preliminary randomization of the features should be carried out.

## Keywords

Classification Algorithms, Granular Computing, Invariants of Matrix Data, Data Processing

---

## 1. Introduction

The classification problem addresses numerous questions in science and technology that involve the use of massive amounts of accumulated information in analyses of new data. These data play the role of incoming information in various models of formalized complex systems. The accumulated information is represented by objects, and the features of objects are vectors, usually in Eucli-

---

\*Independent researcher.

dean space [1] [2] [3].

This representation of objects allows for the use of the powerful technique of linear algebra to study the properties of the corresponding samples. However, this method is based on a sufficiently strong assumption that does not take into account important features of the problem itself. In this problem, real objects are represented by information that describes quantitative images, patterns, and the values of randomly selected features include random errors derived from the measurements or observations. Uncertain data are organically inherent in such tasks relevant to the field of artificial intelligence, in which pragmatic information is processed. A specific task is posed by some expert who, guided by professional expertise, classifies a set of real objects and describes their patterns that are obtained experimentally. Therefore, we here relate the sensory perception of a real object by an expert to the quantitative representation of that object. There clearly cannot be a one-to-one correspondence between the perception of the object and its measurement.

Furthermore, existing methods assume the existence of metric space features and a smooth probability density function for objects of a certain class. However, in a real problem, all objects have similar or identical feature values; thus, the object classes differ only in the probability densities of the features. However, only a generalized function can accurately consider the discontinuous densities of the features. Therefore, each of the existing methods is applied over a restricted area whose boundaries can usually only be established experimentally.

These considerations were taken into account when developing the data matrix invariant method for solving the classification problem [4] [5]. According to this method, a given data matrix is one possible random realization of the set of matrices describing the properties of a sample. This set is here considered to be a set of invariants with respect to the class of objects. The method is based on a concept presented by L. Zadeh. This concept states that the solution of most problems does not require high accuracy because humans perceive only “a trickle of information” from the external world [6]. Moreover, in systems in which complexity exceeds a certain threshold, accuracy and practical sense are almost mutually exclusive.

According to the method, solving the classification problem consists of two stages. In the “training phase” (following [1]), the method calculates a function that uses the simplest and most universal algorithm and establishes a rule according to which the training sample is divided into classes. In the second stage, this function is used to classify new objects.

This article presents a further development of the invariant method for use in the initial stage. A problem in which the properties of a multilevel system are largely independent of the characteristics of the system’s elements is examined here. This type of system can be structured expediently by coarsely “cutting” it into parts that represent information granules, the interactions of which can be described quite simply [7]. The role of such system is played by the aggregating information on the values of the features and the distribution of objects by

classes. Her work, which reduces to the interactions among granules, manifests itself in the frequencies of their appearance.

## 2. Assumptions and Preliminary Observations

The solution is sought in accordance with the following algorithm. The range of values of each feature is divided into an equal number of intervals and the number of each interval, called the index, in which the object feature falls is determined. The objects are then described by a set of indices, and the data matrix is transformed into an index matrix. For each feature, it is possible to find a subset of objects with the same index and these subsets are referred to as granules. The frequency of a certain object's classes can then be calculated in terms of the composition of the granules. The sum of the frequencies of the granules that correspond to individual features is approximately equal to the frequency of occurrence of a certain class object. The maximum of this frequency gives an estimate of the object class for any number of intervals.

The proposed algorithm is conceptually based on the fact that the uncertain quantities are quantities that yield different quantitative information about a single entity. On this basis, the assumption is made here that the given values of the features are random and that values within a certain neighborhood are equally probable. According to this assumption, the probability distribution of features is piecewise constant. Therefore, each feature value of an object can be assigned an index, and for each class, the frequency of the teaching sample is calculated.

The probability of a particular object class consists of a complete group of independent events of index appearance for all features in the same class. Then, the process of classification is reduced to a comparison of the frequencies of different class objects whose feature indices have the frequencies of the corresponding classes. The formula of full probability allows estimating the probability of an object for each class, and its maximum defines an object class.

In general, the compositions of granules (in terms of their feature values and object class) are highly uncertain, and it is advisable to reduce this uncertainty to eliminate possible classification errors. This objective can be partly achieved by increasing the number of intervals to infinity when the granules are composed only of objects with the same feature values.

Since this level is measured by the amount of information entropy, the entropy of the information contained in the data matrix must also be increased. From the properties of entropy, it follows that such a result can be achieved by increasing the number of options in each of the feature values. Practically, this means that it is advisable to reduce the number of equal values of each feature. This conclusion is consistent with the law of necessary diversity [8], which states that an organism's assessment of its environmental impact and adaptation requires the perception of the maximum number of variants of information values. Therefore, this study proposes randomizing the features by increasing their val-

ues by a uniformly distributed random variable that is the same for all features of the object. All of the features then have different meanings, and the training will be asymptotically error-free. An analysis of the solutions for a variety of training sample models shows that this conclusion is valid for any real problem.

The invariant method represents a model of processes in the sensory systems of animals and is intended to carry out class recognition by searching for prototype objects in the information stored in an animal's brain [9]. A single algorithm for implementing the method and model indicates the biological roots of the proposed approach and confirms its applicability.

### 3. Features of the New Multidimensional Data Description System

#### 3.1. Statement of the Problem

The basis of this work is the data matrix invariant method. This method allows the establishment of a function that divides the objects in a training sample into classes. The effectiveness of the method has been demonstrated using 9 databases from a repository [10]. These databases cover a wide range of values in terms of the numbers of objects, features, and classes; the types of features; and the unevenness of the distribution of objects between classes. This article focuses on the limits of the application of the method for the first stage of the solution and the refinement of some features of the method. To this end, we consider the following training problem involving the application of models to sample data.

Let  $\mathbf{X} = \{x_s \mid s \in (1, M)\}$  be sample objects that describe feature vectors  $\mathbf{q}^s = (q_1^s, \dots, q_{Mk}^s)^T$ . Additionally,  $\mathbf{Q}$  is a data matrix;  $\omega_i$  is a list of numbers of objects of class  $i \in (1, Mi)$ ;  $M, Mi$  and  $Mk$  are the number of objects, classes and attributes, respectively, and there are no missing data. In the task, it is necessary to establish a function according to which objects are divided into classes and to assess the influence of various peculiarities of the data on the class recognition accuracy.

The properties of an object are determined by combining the values of its features, and it can naturally be assumed that objects of different classes differ only in the probability distribution of the features. Therefore, each data set is created by the methods of statistical modeling as the union of sets of individual class objects with different feature densities. The values of each feature  $q_k \in \{q_k^1, \dots, q_k^{Mk}\}$  are generated by a mixture  $Mi$  of normal distributions  $N(q_k, \mu_k^i, \sigma_k^i)$ . Here,  $\mu_k^i = k\mu * \mu_0_k^i$  and  $\sigma_k^i = k\sigma * \sigma_0_k^i$  are the sample mean and sample standard deviation of the list of objects  $\omega_i$ , respectively, and  $M, k\mu$  and  $k\sigma$  are model parameters. The values of  $\mu_0_k^i$  and  $\sigma_0_k^i$  indicated in **Table 1**, where  $Mi = 3$  and  $Mk = 4$ , correspond to the Iris database [10]. The distribution of objects by class (5%, 35% and 60%) indicates that the data are “unbalanced”, which complicates the classification problems [11].

Let us note the peculiarities of the models. Data sets are modeled with an arbitrary combination  $M, k\mu \in (0.1, 10)$  and  $k\sigma \in (0.1, 10)$ . By varying the

**Table 1.** Parameters of the initial distribution of features.

Feature	Sample mean			Sample standard deviation		
	$i = 1$	$i = 2$	$i = 3$	$i = 1$	$i = 2$	$i = 3$
$k = 1$	5.1	5.9	6.6	0.35	0.51	0.63
$k = 2$	3.4	2.8	3	0.38	0.31	0.32
$k = 3$	1.5	4.3	5.6	0.17	0.46	0.55
$k = 4$	0.2	1.3	2.1	0.11	0.2	0.27

rounding error, several new samples whose data matrices differ importantly are generated. The variance in several features of one of the classes can be zero. For the features of individual classes and the data set as a whole, the correlation dependence varies widely, from very strong to very weak and from positive to negative.

The sample models are identified by the character string  $a\_b\_c$ , where  $a$  and  $b$  are the values of  $k\mu$  and  $k\sigma$ , respectively, and  $c$  is the rounding parameter. For  $c$ , the values of 0, 1, 2 and 3 correspond to the rounding up of integers, tenths, hundredths and no rounding, respectively. The conclusions obtained in the work are based on an analysis of a large volume of calculated data, the characteristic features of which are illustrated in the figures cited in the paper.

Within the framework of the method, we consider the vector of object features to be an ordered set of feature values to which the segments of the coordinate axes correspond. The object itself is considered to represent a certain point in the multidimensional coordinate system. The totality of such points for all sampling objects determines the feature space. This space is not Euclidean, as is customary in machine learning, because the additional assumption of the existence of the scalar product is not made. Here, we use only the structure of the space, the points of which differ in terms of their numbers and related information.

### 3.2. Multidimensional Intervals

Let us divide the objects into groups in which the values of each feature lie close to one another. For this purpose, we use multidimensional intervals that break the range of values of each feature into an equal number of intervals. For any  $k$ , we then obtain the intervals of values  $q_k \in [q_{k,m}, q_{k,m} + \Delta_k)$  with the step  $\Delta_k = (q_k^{\max} - q_k^{\min}) / (n - 1)$ , where  $n > 1$  is the number of intervals;  $m \in (1, n)$  is the interval number; and  $q_k^{\min}$  and  $q_k^{\max}$  are the minimum and maximum values of the feature, respectively. We call  $m$  the feature's index  $k$  of the object  $s$  if  $q_k^s \in [q_{k,m}, q_{k,m+1})$ .

There are two variants of indexation and of the corresponding function  $\varphi: q_k^s \rightarrow m$ . The first variant realizes the dependence  $m = 1 + W\left(\frac{q_k^s - q_k^{\min}}{\Delta_k}\right)$ , where the function  $W(\cdot)$  calculates the integer part of the number.

The second option assumes that  $q_k^s$  is renumbered and receives a number  $t(s)$

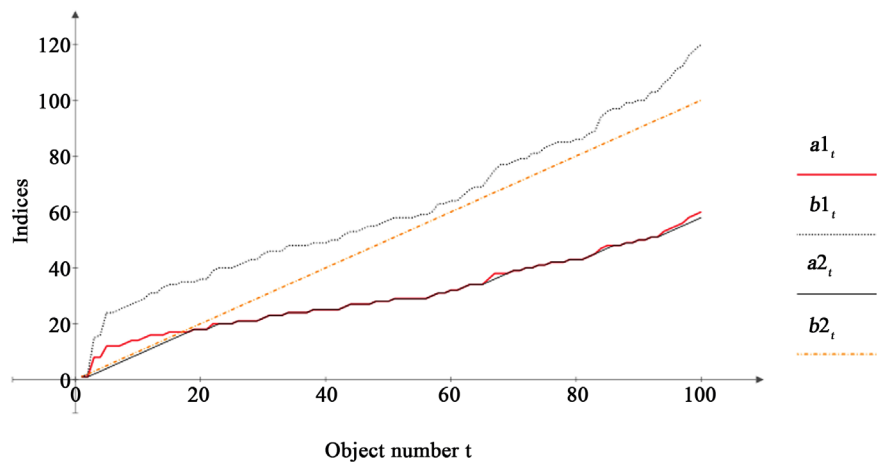
such that the value  $q_k^t$  is arranged in nondecreasing order:  $q_k^1 \leq q_k^2 \leq \dots \leq q_k^M$ . The index  $m = 1$  receives one or several objects (with numbers  $t$ ) for which  $q_k^t \in [q_{k,m}, q_{k,m} + \Delta_k)$ , where  $q_{k,m} = q_k^{\min}$ . If, with a further increase in  $t$ , this relation is not satisfied for some  $t = t_1$ , we then take  $m = m + 1$  and  $q_{k,m} = q_k^{t_1}$  and find the group of objects for which  $m = 2$ . In the same way, we successively determine the indices of all other values of the features.

The indices computed in accordance with options 1 and 2 (hereinafter indices  $i1$  and  $i2$ ) may differ significantly. For example, under  $M = 100$  and  $n = 1000$ , the values of  $i1 \in [1, 1000]$ , and they are described by 100 different numbers, while the values of  $i2 \in [1, 31]$ , and some numbers are repeated many times. The correlation coefficients of the feature vectors and their indices are equal to 1 for  $i1$  and close to 1 for  $i2$ .

The curves in **Figure 1** illustrate the effects of the parameter  $n$  on the index values for the feature vector with  $M = 100$  distributed according to the normal law. (For clarity, the curves are plotted for no decreasing feature values). The curves  $a1$  and  $b1$  correspond to the index  $i1$  when  $n = 70$  and  $120$ , respectively, while similar curves  $a2$  and  $b2$  exist for index  $i2$ . Here and in the other figures, the identifiers of the curves are supplemented with identifiers of the variants of the indices. The graphs show that the indices  $i1$  and  $i2$  almost coincide for  $n < M$  and differ significantly for  $n > M$ .

### 3.3. Granulation of the Given Data

Given the uncertainty of the data, we introduce the assumption that all of the values of the features of the objects fall within a certain interval and include some random errors that are equal. For each indexing option, we find the data matrix mapping  $Q \rightarrow D$ . Here,  $D = (d_{s,k})$  is the index matrix, and  $d_{s,k}$  is the index feature  $k$  of the object  $s$ . The matrix  $D$  has an important feature. Its elements do not characterize the meanings of the features of the objects; instead, these elements indicate the “locations” of their values in the structure of the matrix  $Q$ . If we were to represent some feature value of the object in the form of a



**Figure 1.** The index values for nondecreasing values of the feature vector with  $M = 100$ .

particle, it would be located in the element of the matrix  $\mathbf{D}$  that corresponds to the index for this feature.

The matrix  $\mathbf{D}$  makes it easy to find subsets of objects  $Z_{k,m}$  that have the same values of the ordered pair  $(k, m) = (k, d_k^s)$  for any  $k$  and  $m$ . We call  $Z_{k,m}$  the information granule.

Let  $l_{k,m}$  be the number of objects in the granule  $Z_{k,m}$ , among which  $l_{k,m}^i$  have class  $i$ . Thus, the frequency of objects in class  $i$  in this granule is  $g_{k,m}^i = l_{k,m}^i / l_{k,m}$ . According to the accepted assumption, any object of class  $i$  in the granule  $Z_{k,m}$  has the frequency  $g_{k,m}^i$  and therefore the sample probability  $p(s \in \omega_i | d_k^s = m) = g_{k,m}^i$ .

The event occurrence of vector  $\mathbf{d}^s$  for objects of class  $i$  consists of the complete group of independent events  $d_1^s, \dots, d_{Mk}^s$ . It follows from the total probability formula that the probability estimate for this event is

$$p(s \in \omega_i | \mathbf{d} = \mathbf{d}^s) = \frac{1}{Mk} \sum_{k=1}^{Mk} g_{k,m}^i = \bar{g}_{k,m}^i \quad (1)$$

Here,  $\bar{g}_{k,m}^i$  is the average frequency  $g_{k,m}^i$  for all  $k$ . The calculated class of object  $s$  corresponds to the maximum of this value for a given  $n$ :

$$I_n = \arg \max_{1 \leq i \leq Mi} p(s \in \omega_i | \mathbf{d} = \mathbf{d}^s) \quad (2)$$

This formula establishes a rule for recognizing an object class as a result of training using the sample data. According to (2), the object class depends on the frequency  $g_{k,m}^i$ , which serves as an objective measure of the membership of any object from the set  $X$  to a certain class based on separate features. The evaluation of the class is robust because the grouping of objects is performed in its calculation. This operation neutralizes the influence of sharply differing feature values.

The accuracy of solving the training problem is determined by examining the fulfillment of the relation  $I_n(s) = i^s$  for each sample object. It is estimated by the frequency  $\gamma_n$ , which is equal to the ratio of the total number of errors to the length of the sample.

### 3.4. Influence of the Parameter $n$

We are interested in the properties of the sequence  $\{\gamma_n\}$  where  $n = 2, 3, \dots$ . For some  $k$ , let the granule  $Z_{k,m}$  contain an object  $s$  of class  $i$ , as well as other objects, including the object  $w$ . Therefore, in any scheme, indexing is performed following the inequality  $0 \leq |q_k^w - q_k^s| \leq \Delta_k$ .

Consider the limiting case of  $n \rightarrow \infty$  corresponding to the step  $\Delta_k \rightarrow 0$ . It follows from the above relation that  $q_k^w = q_k^s$ . Since all of the values of the feature are different in statistical modeling and in the absence of rounding, this equality is satisfied only when  $w = s$ . The granule  $Z_{k,m}$  then consists of a single object  $s$  having class  $i$ , and the frequency  $g_{k,m}^i = 1$  for any  $k \in (1, Mk)$ . Hence,  $I_n(s) \rightarrow i^s$  for all  $s \in (1, M)$ , and the sequence  $\{\gamma_n\}$  converges in probability to zero.

Thus, Formulas (1) and (2) define a function that asymptotically error-free

calculates the class of any object of the training sample.

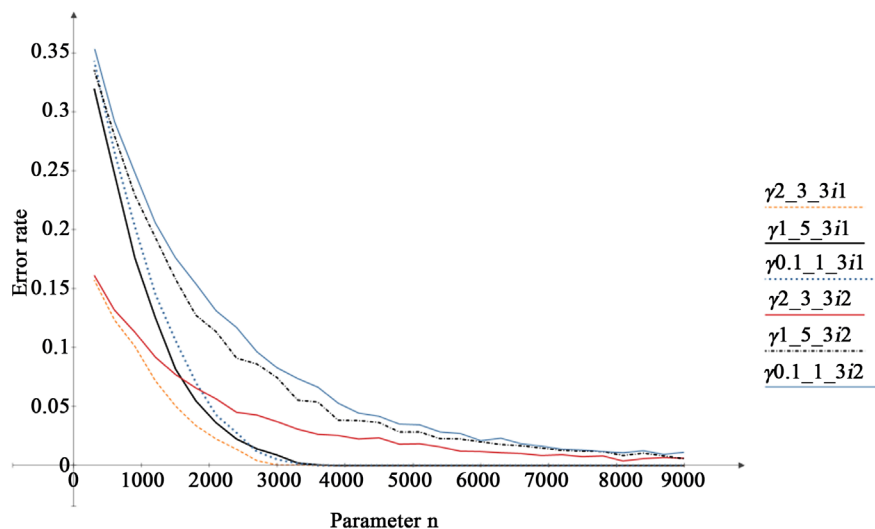
The results of calculations of the function  $f(n) = \gamma_n$  for several samples at a value of  $n = 3000$  are illustrated by the curves shown in **Figure 2**. For all of the curves, a notation system is adopted in which the symbol  $\gamma$  is followed by a string of characters that act as the identifier of the sample, and the indexing option  $i1$  or  $i2$  is then indicated. The graphs depict the sequence convergence process  $\{\gamma_n\}$ . They show that the convergence rate for index  $i1$  is higher than that for  $i2$ .

The calculations show that the correlation coefficients of the features of individual classes and the corresponding indices  $i1$  and  $i2$  do not differ significantly.

### 3.5. Randomization of the Data

The question of the limitations of the given data in solving the training problem using the method considered naturally arises. In connection, we note that the curves in **Figure 2** are plotted for the samples in which all of the features have different values, as obtained directly by statistical modeling (for  $c = 3$ ). In real machine learning tasks, objects often have repeated, identical values. This situation is not uncommon in the case of quantitative traits and is common in problems in which traits are not examined quantitatively, but are represented by, for example, nominal and ordinal features. Here, this situation is modeled by rounding the feature values because the number of different feature values in the matrix  $Q$  can be significantly reduced by rounding. For example, if the value of  $c$  is decreased from 3 to 0 in sample 1\_0.1\_3, all of the values of  $q_2$  will be equal to 3.

The effect of rounding on the accuracy of training occurs via following mechanism. Consider a granule  $G_{k,m}$  containing an object  $s$  of class  $i$  for some  $k$  in the case in which  $n \rightarrow \infty$ . Suppose that there is a set of objects  $\{w_1, w_2, \dots\}$  of different classes in which  $q_k^{w_1} = q_k^{w_2} = \dots = q_k^s$ . For any other  $k$ , a similar ratio



**Figure 2.** Influence of the parameter  $n$  on the accuracy of training.



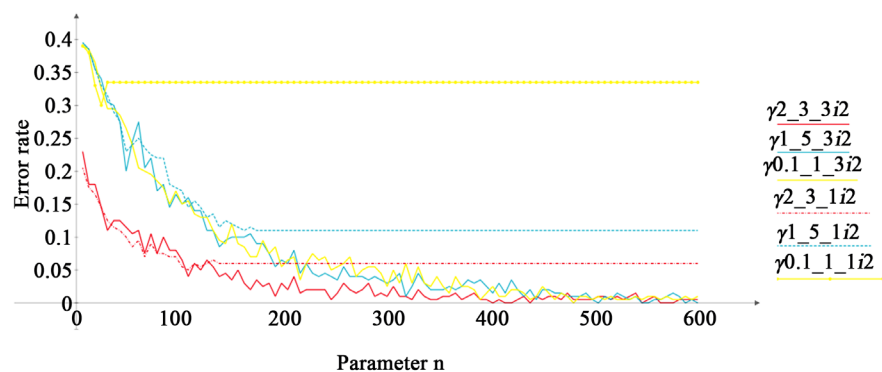
can be constructed. However, it will concern an analogous set made up entirely or partially of other objects because two objects with the same characteristics cannot exist in the sample. All of the summands that determine the value of  $\bar{g}_{k,m}^i$  for the object  $s$  will be greater than zero. Since  $M \ll n$ , most of the analogous summands for the objects of the indicated sets will differ from zero only for some values of  $k$  and  $i$ . Therefore, given repeated values of the attributes, training errors occur.

**Figure 3** shows the results of calculations for the three samples before and after rounding to the nearest tenth ( $c = 3$  and 1) for  $M = 200$  and index  $i2$ . This figure shows that rounding leads to significant changes. A sequence  $\{\gamma_n\}$  does not converge in probability to zero; for sufficiently large values of  $n$ , there are portions with constant values  $\gamma_n$ .

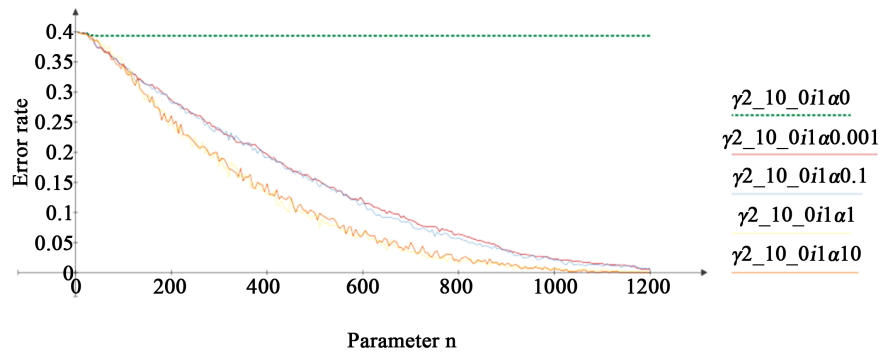
This situation is explained by the fact that rounding changes the level of diversity in the data  $\beta$ , which is equal to the average number of different feature values in the sample. For example, for the sample 2\_3\_2 with  $M = 1000$ , a sequence of values  $c = \{3, 2, 1, 0\}$  corresponds to  $\beta = \{1000, 478, 87, 11\}$ . According to the theory of K. Shannon, the amount of information per one feature value increases proportionally to  $\log_2(1000)/\log_2(\beta)$ . In this case, this volume changes in the ratio 1:1.12:1.55:2.88. The corresponding increase in “information overload” leads to a situation in which, for some objects, the values  $\bar{g}_{k,m}^i$  will be the same for different  $i$  given a sufficiently large  $n$ . Formula (2) will then not be able to correctly predict the class of objects.

To avoid such errors, we use the idea of invariants and move from the data matrix to its invariant, in which features do not have repeating values. The invariant of the data matrix is found by randomization according to the assumption that, for all  $k \in (1, Mk)$  and  $s \in (1, M)$ , we can replace  $q_k^s$  by  $q_k^s + \alpha v_s$ , where  $v$  is a random variable uniformly distributed on the segment  $[0, 1]$  and  $\alpha \geq 0$  is a constant. Since the average values of each feature increase by  $\alpha/2$  with this transformation, the correlation dependence of the features will not change.

The influence of the parameter  $\alpha$  on the training accuracy is shown in the curves given in **Figure 4**. The curves represent the results of calculations for the sample 2\_10\_0 with  $M = 1000$ , the index  $i1$  and the five values  $\alpha = 0, 0.001,$



**Figure 3.** Effect of rounding on the accuracy of training.



**Figure 4.** The impact of randomization on the accuracy of training.

0.1, 1 and 10. The curves show that the randomization effect occurs for any  $\alpha > 0$ . Since the values of the features are located on the segment  $[-7, 29]$ , randomization results in significant relative changes for some of these values. However, for all values of  $\alpha$ , high-quality training is achieved; for these data sets, the influence of the parameter  $\alpha$  is reflected only in the rate of convergence.

Thus, due to randomization, we cannot completely neutralize the effects of low levels of diversity and achieve error-free training.

### 3.6. Limitations of the Given Data

In the initial stage of sample modeling, the objects of each class differed in the distribution of their features. Rounding violated this provision as columns of the data matrix ceased to represent the calculated distribution. Nevertheless, for all of the samples, error-free training was achieved. Below, we consider the question of whether error-free training will be achieved for any sample.

Earlier, we examined samples that have equal values for some of the features and showed that the application of randomization enables the achievement of error-free training. It follows from computational experience that the case in which a relatively large number of objects have the same feature values requires special consideration.

Let each value  $p$  of the feature occur  $r_p$  times and  $p \in (1, l), l \leq M$ . In this case, the number of variants of the placement of the feature values is

$$K = \frac{M!}{r_1! * r_2! * \dots * r_l!}$$

where  $\sum_{p=1}^l r_p = M$  [12]. For  $l = M$ , we obtain the limiting case in which all of the feature values differ from each other, for which  $r_1 = r_2 = \dots = r_M = 1$  and  $K_1 = M!$ .

If only  $v$  of  $M$  objects have different feature values, then  $r_1 = v$  and  $r_{v+1} = r_{v+2} = \dots = r_M = 1$ . The number of variants is then equal to  $K_2 = \frac{M!}{v!}$ , and the attribute entropy decreases from  $\log_2(K_1)$  to  $\log_2(K_2)$ . In another limiting case in which all of the feature values are the same (as in the sample 1\_0.1\_0)), we obtain  $v = M$  and  $\log_2(K_2) = 0$ . Here, the feature carries no

information and is not actually taken into account, although it is “claimed” to be one of the features. It is thus a “degenerate” feature.

From the above dependences, it follows that the level of decrease in the entropy of the feature and the information it contains depends on the ratio of  $M$  to  $v$ . If  $M = 100$ , then the maximum value of the entropy is  $\log_2(100!) = 363$  bits due to the decrease in the repeated values of the features. Furthermore, for the sequence of values  $v = \{70, 90, 99\}$ , there will accordingly be 133, 45, 4.6 bits.

As noted in the Introduction, the training problem, which is a stage of classification, is formulated by an expert. Only this expert can decide whether a matrix carrying a limited amount of information can serve as a data matrix. For us, only the case  $v = M$  is obvious. In this case, the “degenerate” feature will receive random information during the randomization stage, possibly distorting the essence of the problem.

A similar situation occurs if the sample variance for the entire sample or one class (for example, sample 10\_0.1\_3) is close to zero. Most of the values of this attribute will then be concentrated near the mathematical expectation, and it may turn out that all of the other values are represented by an extremely limited number of indices. Note also that an uneven distribution of objects by class can enhance the effect of reducing the volume of information. In such cases, the sequence  $\{\gamma_n\}$  cannot converge to zero. If the resulting error is significant, the scheme for computing the invariants of the data matrix should then be changed by first converting the values of the corresponding feature. A similar transformation for the Adult database led to error-free training [5].

For real datasets, we can assume that the invariant method leads to practically error-free training.

## 4. Conclusions

This paper presents a solution to the problem of training sample models by employing the data matrix invariant method, which was developed to solve the classification problem. This solution is limited to finding a function according to that the objects in a training sample are divided into classes, studying the influence of various data features on the accuracy of training.

The key idea of the method is structuring the information contained in the problem by introducing multidimensional intervals that allow us to roughly break up the given dataset into its component parts and to compute the information granules for each feature. The granules possess an important property in that, given an infinite number of interval; each granule contains only the objects with the same feature value.

The principle of the scheme of structuring information is reduced to dividing the whole, which is represented by multilevel data, into parts such that one can reproduce the whole from the lower levels to calculate the frequencies of these parts. This “whole” consistently represents the summary information of the sample, the classes, the objects, the features and the indices of the features.

Therefore, multidimensional intervals allow us to compare objects in a probability space, and they play the role of the concept of distance in determining the positions of objects in Euclidean space. The advantage is that the evaluation of the probability of an object is incomparably closer to solving a problem than evaluating the position of an object in a metric space.

In this paper, it is shown that the data matrix invariant method yields practically error-free training for any data matrix on the basis of a simple and universal algorithm. This method has independent importance because it offers a new way of analyzing multidimensional data that do not rely on the concept of distance between objects.

### Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

### References

- [1] Bishop, C. (2006) Pattern Recognition and Machine Learning. Springer, Berlin, 738.
- [2] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer, Berlin, 764.  
<https://doi.org/10.1007/978-0-387-84858-7>
- [3] Murphy, K. (2012) Machine Learning. A Probabilistic Perspective. MIT Press, Cambridge, Massachusetts, London, 1098.
- [4] Shats, V.N. (2016) Invariants of Matrix Data in the Classification Problem. *Stochastic Optimization in Informatics*, **12**, 17-32.
- [5] Shats, V.N. (2017) Classification Based on Invariants of the Data Matrix. *Journal of Intelligent Learning Systems and Applications*, **9**, 35-46.  
<https://doi.org/10.4236/jilsa.2017.93004>
- [6] Zadeh, L. (1979) Fuzzy Sets and Information Granularity. In: Gupta, N., Ragade, R. and Yager, R., Eds., *Advances in Fuzzy Set Theory and Applications*, World Science Publishing, Amsterdam, 3-18.
- [7] Yao, J., Vasiliacos, V. and Pedrycz, W. (2013) Granular Computing: Perspective and Challenges. *IEEE Transactions on Cybernetics*, **43**, 1977-1989.  
<https://doi.org/10.1109/TSMCC.2012.2236648>
- [8] Ashby, W.R. (1957) An Introduction to Cybernetics. 2nd Edition, Chapman and Hall, London, 294.
- [9] Shats, V.N. (2018) The Classification of Objects Based on a Model of Perception. In: Kryzhanovsky, B., Dunin-Barkowski, W. and Redko, V., Eds., *Advances in Neural Computation, Machine Learning, and Cognitive Research, Studies in Computational Intelligence*, Springer, Cham, 3-8. [https://doi.org/10.1007/978-3-319-66604-4\\_1](https://doi.org/10.1007/978-3-319-66604-4_1)
- [10] Asuncion, A. and Newman, D.J. (2007) UCI Machine Learning Repository. Irvine University of California, Irvine.
- [11] Lopez, V., Fernandez, A., Garcia, S., Palade V. and Herrera F. (2013) An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences*, **250**, 113-141.  
<https://doi.org/10.1016/j.ins.2013.07.007>
- [12] Anderson, J.A. (2003) Discrete Mathematics with Combinatorics. Prentice Hall, Upper Saddle River, NJ, 784.