ISSN 1990-4789, Journal of Applied and Industrial Mathematics, 2013, Vol. 7, No. 2, pp. 275–286. © Pleiades Publishing, Ltd., 2013.
Original Russian Text © N.G. Zagoruiko, I.A. Borisova, O.A. Kutnenko, V.V. Dyubanov, 2013, published in Sibirskii Zhurnal Industrial'noi Matematiki, 2013, Vol. XVI, No. 1, pp. 29–41.

A Construction of a Compressed Description of Data Using a Function of Rival Similarity

N. G. Zagoruiko[1,2,3*], I. A. Borisova[1], O. A. Kutnenko[1***], and V. V. Dyubanov[1****]**

[1] *Sobolev Institute of Mathematics, pr. Akad. Koptyuga 4, Novosibirsk, 630090 Russia*

[2] *Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090 Russia*

[3] *Design Technological Institute of Digital Techniques, ul. Akad. Rzhanova 6, Novosibirsk, 630090 Russia*

Received December 6, 2012

Abstract—We justify the claim that all kinds of data mining aim at constructing compressed and simplified descriptions of information. We propose passing from the binary absolute measure of similarity between objects to a ternary relative measure: a function of rival similarity (FRiS-function). Its use enables us to obtain a quantitative estimate for the compactness of data and construct new, more effective cognitive analysis methods. We present examples of solutions to various model and real problems by the new methods.

DOI: 10.1134/S199047891302018X

Keywords: *data mining, function of rival similarity, compactness, pattern recognition, object censoring, feature selection*

INTRODUCTION

The most important problem in cognitive data mining is to systematize and structure the results of observations or experiments in a simplified and compressed form accessible for understanding and further use. We can compress information, for instance, by decreasing the number of objects considered and replacing the entire sample with a small number of its typical (standard) representatives. Another version of data compression decreases the number of characteristic features. In this case, we choose from the set of observable features those most relevant to the problem in question. Using these approaches, we can pass from initial data of arbitrarily large size a tangible compressed description. The main requirement imposed on this description is the condition that the dependencies important for the problem in question must be retained.

A compactness conjecture helps us to ensure that pattern recognition meets this requirement. In the framework of the conjecture, we posit the existence of a feature space in which all objects of the sample are divided into easy to distinguish compact groups (clusters) of similar objects. Similarity within the groups enables us to replace the set of objects in each group by a standard object, while a reasonable choice of the subset of features ensures that each group is homogeneous with respect to the initial classification, the information about which we must keep.

It currently remains an open question how to estimate the compactness of data. The definition of compactness in [1] rests on the ratio of the number of "interior" and "boundary" points representing the objects of the patterns in the feature space. Instead of one quantitative characteristic, [2] calculates a "compactness profile" reflecting the dependence of the number objects of "its own" pattern in a local neighborhood of each object of the sample on the radius of this neighborhood. The compactness of clusters is estimated in [3] by averaging the squared distance from the objects to the centers of their clusters.

* E-mail: zag@math.nsc.ru
** E-mail: biamia@mail.ru
*** E-mail: olga@math.nsc.ru
**** E-mail: vladimir.dyubanov@gmail.com

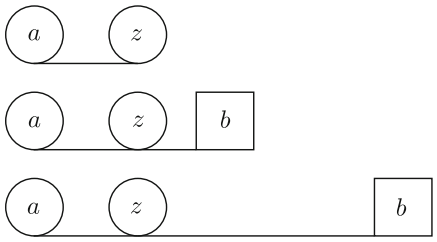

Fig. 1. Rival similarity of objects a and z

In this article, we present a new method for obtaining a quantitative estimate for the compactness of patterns based on a measure of similarity between the objects, which we call the *function of rival similarity*, FRiS-*function* [4]. The use the FRiS-function enables us to develop new cognitive analysis methods aimed at data compression, making them more universal, noise immune, and insensitive to the distribution of patterns and the ratio of the numbers M and N of objects and features. Furthermore, the orientation at the compactness conjecture ensures good agreement with the requirements of cognitive computations regarding the transparency of the solution process and the interpretability of the results.

Following the description of the proposed FRiS-method for estimating compactness, we briefly survey the algorithms based on it for solving known and some new data mining problems and present the results of solving some model and applied problems.

1. THE FUNCTION OF RIVAL SIMILARITY

We often use standard objects to measure the weight, length, resistance, and other characteristics of objects. The result of a measurement is determined solely by the properties of the measured and standard objects, but is independent of the properties of other objects. For this reason, the result is an absolute quantity. However, objects can be described by characteristics like "similar/dissimilar", "near/far", "good/evil", and so forth. There are no standards for this kind of features. Two objects with distinct properties can be regarded as "similar" or "dissimilar", "near" or "far" depending on the properties of other objects. For instance, in Fig. 1 the distance between objects a and z remains the same, but the answers to the question whether they are sufficiently close to each other to be gathered in one class are different in all three cases.

Adequate similarity measures must depend on the particular features of the rival surroundings of z. In the attempted recognition whether an object z belongs to one of two patterns A and B using heuristic recognition algorithms based on the similarity of objects, it is important to know not only the distance $r(z, A)$ from z to the pattern A, but also the distance $r(z, B)$ from z to the rival pattern B. Consequently, in pattern recognition similarity is a relative category rather than absolute. Note that depending on the specifics of the problem in question the distance $r(z, A)$ from an object z to an pattern A can be calculated in different ways. We may use either the distance $r(z, a)$ to the nearest object a of A, or the average distance to all its objects, or the average distance to its k nearest objects, or the distance to its center of mass, and so forth. For instance, in the method of k *nearest neighbors* (kNN) a new object z is recognized as an object of the pattern A whenever the average distance $\bar{r}(z, A(k))$ from z to k nearest objects of this pattern is not only small, but is less than the average distance $\bar{r}(z, B(k))$ to k nearest objects of the rival pattern B. In this algorithm, we estimate similarity on the order scale.

The RELIEF algorithm [5] uses a more complicated similarity measure. In order to define the similarity of an object z to an object a in competition with an object b, it uses a quantity accounting for the normalized difference of the distances $r(z, a)$ and $r(z, b)$:

$$W(z, a|b) = \frac{r(z, b) - r(z, a)}{r_{\max} - r_{\min}}.$$

Here $r_{\min}$ and $r_{\max}$ are the minimal and maximal distances between all pairs of objects of the analyzed set.

In order to estimate the "silhouette width," [6] measures the average distance $\bar{r}(z, A(M))$ from an object $z \in A$ to all M objects of A and the distance $r(z, b)$ from z to the nearest object $b \notin A$. The measure of *similarity* of z to the objects of A is taken to be

$$S(z, A|b) = \frac{\bar{r}(z, A(M)) - r(z, b)}{\max\{\bar{r}(z, A(M)), r(z, b)\}}.$$

We propose to use a FRiS-function to calculate the similarity of objects. This is a ternary relative measure estimating the similarity of an object z to an object a in competition with an object b:

$$F(z, a|b) = \frac{r(z, b) - r(z, a)}{r(z, b) + r(z, a)}. \tag{1}$$

As z moves from a to b, we can speak firstly about great similarity of z to a; then about their moderate similarity; then about the onset of the same similarity, equal to zero, to both a and b. As z further moves toward b, first moderate, and then large differences between z and a appear. The coincidence of z and b means the maximal distinction of z from a, which corresponds to the similarity of z to a equal to -1.

Both the distance r between objects and the similarity F between them are independent of the location of the origin of the coordinate system, the rotations of the coordinate axes, and the simultaneous multiplication of their values by the same quantity. However, independent changes in the scale of different coordinates change the contribution of separate peculiarities to both distance and similarity estimates. Thus, the similarity of objects depends on the weights of different features. Changing these weights, we can emphasize the similarity or distinction of the specified objects, which is usually done when choosing relevant features and constructing decision rules in pattern recognition.

Similarity in the order scale in the kNN method answers the question the objects of which pattern object z is most similar to. Rival similarity measured using the FRiS-function answers this question, and moreover, the question what the absolute value is of the similarity of z to the objects of A in competition with the objects of B. It turned out that the additional information which the absolute scale yields in comparison with the order scale enables us to substantially improve data mining methods.

We define the rival similarity of objects to patterns following the same principle as the rival similarity of objects to objects:

$$F(z, A|B) = \frac{r(z, B) - r(z, A)}{r(z, B) + r(z, A)}.$$

In the case of normally distributed patterns with the same covariance matrices, we can calculate the rival similarity of an object to these patterns via the similarity to their average. However, if the patterns have very complicated structure then, for calculating the FRiS-function, we can only inspect the local neighborhood (the nearest neighbors) of the object it is calculated for.

Solving the recognition problem on assuming that it is possible to estimate the variances d_A and d_B of the distributions of the rival patterns A and B, we have to use the normalized distances from the object z to the patterns A and B. The resulting normalized function of rival similarity is

$$F_d(z, A|B) = \frac{d_A r(z, B) - d_B r(z, A)}{d_A r(z, B) + d_B r(z, A)}.$$

One of the methods for extracting the specific peculiarities of data in the recognition problem is to pass to their compressed description using sets of standard representatives of each pattern which retains the basic dependencies necessary for good recognition of both the objects of the initial sample and new objects. Henceforth, we refer to these standard objects as *stolps*. The more complicated is the structure of the patterns and the stronger they intersect , the more stolps we will need to describe the data. If we manage to construct this description of data and pass from A and B to stolp sets S_A and S_B for these patterns then we can calculate the rival similarity of the object z to the pattern A in competition with B as $F(z, S_A|S_B)$. Calculating rival similarity using the compressed description rather than the whole sample, we can adapt this measure to the specific features of the problem in question.

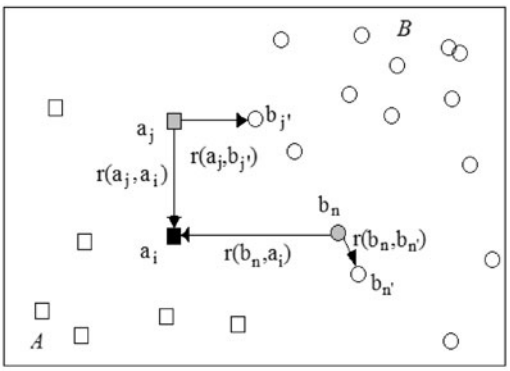

Fig. 2. Estimates for the defensibility and tolerance of the object $a_i \in A$

2. THE CHOICE OF STANDARD OBJECTS. ALGORITHM FRiS-Stolp

In order to construct a compressed description of data as a system of stolps, we use the FRiS-Stolp algorithm [7]. It works for every ratio of the number of objects to the number of features, as well as for arbitrary distributions of the patterns.

As the stolps we choose the objects with high values of the two properties: *defensibility* with respect to the objects of its own pattern and *tolerance* with respect to the objects of other patterns. The higher is the defensibility of a stolp, the fewer errors of the first kind (missing the goal) will occur. The higher is the tolerance of a stolp, the fewer errors of the second kind (false alarm) will occur. We regard a collection of stolps as sufficient to describe the sample whenever the similarity F of all objects of the training sample to the nearest its own stolps in competition with the nearest objects of the other patterns exceeds the threshold value F^*, for instance, $F^* = 0$.

Let us use Fig. 2 to illustrate a method for estimating the tolerance and defensibility of an object on an example of the recognition problem for two patterns $A = \{a_1, \ldots, a_{M_A}\}$ and $B = \{b_1, \ldots, b_{M_B}\}$ expressed respectively as tuples of M_A and M_B objects of the training sample.

Verify whether the object a_i defends well the objects a_j of pattern A, for $j = 1, \ldots, M_A$. For the object a_j, define the distances $r(a_j, a_i)$ and $r(a_j, b_{j'})$, where $b_{j'} \in B$ is the nearest neighbor of a_j; thus,

$$j' = \arg \min_{m=1,\ldots,M_B} r(a_j, b_m).$$

Using (1), we obtain the value $F(a_j, a_i|b_{j'})$ of the similarity function of a_j to $a_i \in A$ in competition with $b_{j'} \in B$ (see Fig. 2). Select those objects $a_j \in A$ for $j = 1, \ldots, M_A$ whose similarity to a_i is at least the specified threshold F^*; thus, $F_j^+ = F(a_j, a_i|b_{j'}) - F^* \geq 0$. These objects are defended well by a_i. We obtain the estimate D for the defensibility of a_i:

$$D(a_i) = \sum_{j=1}^{M_A} F_j^+|_{F_j^+ \geq 0}.$$

Now estimate the tolerance of a_i which measures the dissimilarity to a_i of the objects of pattern B. For every $b_n \in B$ with $n = 1, \ldots, M_B$, calculate the distances $r(b_n, a_i)$ and $r(b_n, b_{n'})$, where $b_{n'} \in B$ is the nearest neighbor of b_n. Using (1), find the similarity $F(b_n, b_{n'}|a_i)$ of b_n to $b_{n'}$ in competition with a_i (see. Fig. 2). Select the objects of pattern B with $F_n^- = F(b_n, b_{n'}|a_i) - F^* < 0$ for $n \in \{1, \ldots, M_B\}$. These objects are more similar to a_i than to the nearest objects of their own pattern, which negatively affects the estimate for a_i. We obtain the following estimate T for the "intolerance" of object a_i:

$$T(a_i) = \sum_{n=1}^{M_B} F_n^-|_{F_n^- < 0}.$$

We estimate the quality of object a_i in the role of a stolp of A as $S(a_i) = D(a_i) + T(a_i)$.

Note some specific characteristics of the FRiS-Stolp algorithm. Independently of the distribution of the training sample, we choose as stolps the objects lying near the centers of local clusters and defending as many objects as possible with a specified reliability. For normal distributions we choose as stolps first of all the objects nearest to the expected value. Consequently, as the distribution approaches a normal one, the solution to the problem of constructing decision functions tends to the statistically optimal. If the distributions are polymodal and the patterns are linearly inseparable then the stolps lie at the centers of the modes.

We can use the compressed description of the patterns as stolp sets to recognize new objects. The process of recognition consists in the following:

(1) Find the distances from the control object z to two nearest stolps belonging to different patterns.

(2) The object z will belong to the pattern whose stolp is closer.

(3) Using the distances, evaluate the function of rival similarity F of the object to the patterns. The quantity F enables us to judge the reliability of the decision made.

If the number K of the patterns is greater than two then in constructing the stolps for the pattern A_k with $k \in \{1, \ldots, K\}$, we gather the objects of all remaining patterns into one virtual pattern

$$B_k = \bigcup_{i=1,\ldots,K,\, i \neq k} A_i.$$

3. FRiS-COMPACTNESS

Given an object $a \in A$, the measure of rival similarity of this object to its own pattern in competition with an pattern B shows how much this object is similar to its own pattern and dissimilar to B. If this quantity is positive for all objects of A then we may regard this pattern as compact since this situation agrees well with the intuitive representation of compactness as the similarity of objects in the pattern and their dissimilarity to the objects of the competing pattern. Thus, calculating the average value of the FRiS-function over all objects of A, we can estimate the compactness of this pattern. Furthermore, if we calculate the FRiS-function basing on stolps then this compactness estimate automatically adapts to the specific characteristics of the data.

In the case of two patterns $A = \{a_1, \ldots, a_{M_A}\}$ and $B = \{b_1, \ldots, b_{M_B}\}$ we propose the following version of the compactness estimate.

(1) Using the FRiS-Stolp algorithm, construct c stolps of the patterns A and B; hence, $c = c_A + c_B$, where c_A and c_B are the numbers of stolps of the patterns A and B respectively.

(2) For every element $a_i \in A$, estimate the similarity to its own nearest stolp $s_A(a_i)$ in competition with the nearest stolp $s_B(a_i)$ of B. Then calculate the FRiS-compactness of A in competition with the pattern B as

$$C_{A|B} = \frac{1}{c_A M_A} \left(\sum_{i=1}^{M_A} F(a_i, s_A(a_i) | s_B(a_i)) - c_A \right). \tag{2}$$

(3) Similarly calculate the FRiS-compactness $C_{B|A}$ of B in competition with A.

(4) Obtain the compactness estimate for A and B as the average of the values of $C_{A|B}$ and $C_{B|A}$.

Observe that the number c_A of clusters of A depends on the structure of the distribution of objects and the threshold F^*: as F^* grows, the number of clusters and the accuracy of description of the distribution increase, but so does the complexity of the description, that is, the factor $1/c_A$ is the penalty for the structural complexity of the pattern.

If the number of patterns K is greater than two then to estimate the compactness of A_k for$k \in \{1, \ldots, K\}$, we gather the objects of all remaining patterns into one virtual pattern B_k. With the compactness estimates $C_{A_k|B_k}$ for $k = 1, \ldots, K$ of all patterns already at hand, we can obtain the total estimate for their compactness in this feature space as their mean:

$$C = \frac{1}{K} \sum_{k=1}^{K} C_{A_k|B_k}.$$

If we seek to maximize the FRiS-compactness of the least compact pattern then we have to use the average estimate

$$C = \sqrt[K]{\prod_{k=1}^{K} C_{A_k|B_k}}. \tag{3}$$

Our experiments with these two FRiS-compactness criteria showed the superiority of the second of them (see [8]).

The FRiS-compactness of patterns equals 1 whenever all objects of each pattern are mapped to their own separate points. If, at every nonempty point of the space, there is one object of at least two patterns then the distance to the nearest its own object is more than 0, while the distance to the nearest foreign object equals 0, and the FRiS-compactness of the patterns equals -1. All remaining distributions yield the values of FRiS-compactness in the range from -1 to 1.

4. THE CHOICE OF STOLPS FOR A SET OF UNCLASSIFIED OBJECTS. THE FRiS-TAX ALGORITHM

We showed above how to obtain a compressed description of the structure of data by selecting a set of typical representatives from the sample for the recognition problem in which for each object we know by name the class it belongs to.

Let us show how we can construct a collection of stolps for the taxonomy problem, in which the names of the classes for all objects of the sample are unknown. To store the information on the structure of the initial data, we place stolps at the centers of the zones of local clusters of objects so that the objects in each cluster are more similar to its stolp than to all other stolps.

Since here the terms "similar/dissimilar" appear again, it is natural to suppose that a function of rival similarity will enable us to solve this problem. It is only necessary to redefine this function for the case of unclassified data, when beforehand it is unknown which objects of the sample are "its own", and which are "rival". We assume that they all belong to an pattern A. Then, as the distance from an arbitrary object a to its pattern we take the distance $r(a, A)$ to the pattern A. In order to create a competitive situation we introduce a virtual pattern B, the distance to which from every object of the initial sample (that is, the pattern A) is fixed at r^*. The resulting FRiS-function for the taxonomy problem is

$$F(a, A|B) = (r^* - r(a, A))/(r^* + r(a, A)).$$

It is rather easy to imagine how the virtual pattern B looks. To the n-dimensional space in which the sample A is described, we have to add the $(n + 1)$st coordinate whose value for the objects in A is set to zero. Then we can take as B the set points coinciding with the objects of A in the space of the first n features for which the value of the $(n + 1)$st feature equals r^*.

The FRiS-Cluster algorithm, which chooses a system of stolps for an unclassified sample, is a part of the FRiS-Tax algorithm [9] for solving the taxonomy problem. On the first stage of FRiS-Tax, we run a procedure for choosing stolps and dividing the objects of an unclassified sample into linearly separable clusters. This partition can already be regarded as the result whenever the expect is satisfied with its quality. Otherwise, we forward the resulting clusterization to the second stage, the FRiS-Class algorithm, which analyzes the situation at the boundaries of clusters and combines some neighboring clusters into the classes of arbitrary form (taxons) which are not necessarily linearly separable.

Fig. 3 depicts the results of solving the problem of classifying the metal alloys according to their chemical composition. The sample consisted of the X-ray spectra of 160 sample alloys divided into 5 groups by their chemical composition. Every spectrum was represented as a vector of dimension 1024. We partitioned the sample into classes, whose number varied between 2 and 15, using several available taxonomy algorithms. The quality of taxonomy was estimated by the chemical homogeneity of the clusters in terms of entropy.

We compared five algorithms: FRiS-Cluster; FRiS-Tax; the Forel algorithm rolling the objects into clusters of spherical shape [10]; the Scat algorithm [10] constructing from the clusters created by the Forel algorithm taxons of a more complicated form; and the most popular k-means algorithm [3, 11].

These results show that the taxonomy constructed by the FRiS-Tax algorithm is better than those obtained by the other algorithms used for the comparison. Moreover, this algorithm enables us to automatically determine the preferable number of clusters from the first local maximum of FRiS-compactness.

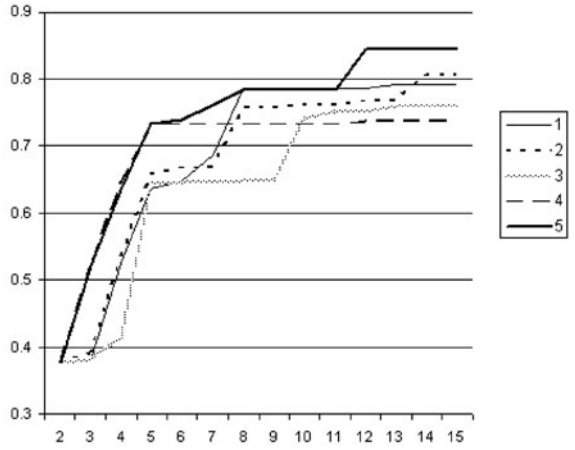

Fig. 3. The dependence of the quality of taxonomy on the number of taxons for algorithms FRiS-Cluster (1), k-means (2), Forel (3), Scat (4), and FRiS-Tax (5)

5. GENERALIZED CLASSIFICATION. THE FRiS-TDR ALGORITHM

We introduced above the function of rival similarity for classified and unclassified samples and considered the algorithms for constructing a concise description of data using typical representatives. Proceed now to a more general case when the sample includes both classified and unclassified objects. Mixed samples of this type appear in the semi-supervised learning problem.

This problem is an intermediate step between the taxonomy and recognition problems. In solving this problem, to construct decision rules we use both objects with specified names of the classes and objects for which the names of the classes are unavailable. If the number of unclassified objects in the sample is considerably greater than the number of classified objects then we use the information on the properties of unclassified objects, which enables us to better understand the properties of the sample and construct a decision rule of higher quality in comparison with what would be constructed basing only on the classified objects in the framework of the recognition problem.

Consider how the technique for calculating the function of rival similarity for the objects to "their own" classes will change as we pass to a mixed sample V_{mix} containing both classified and unclassified objects. In the case of the recognition problem of two patterns A and B, we can divide all objects of the mixed sample into the three groups: $V_{\text{mix}} = A \cup B \cup U$. The group U consists of the objects for which the name of the pattern is unknown. Let U_A (U_B) denote the set of objects U belonging to the pattern A (B). Hence, $U = U_A \cup U_B$. The function of rival similarity is calculated as

$$F(z, A|B, U) = \big(r(z, B \cup U_B) - r(z, A \cup U_A)\big)/\big(r(z, B \cup U_B) + r(z, A \cup U_A)\big).$$

The absence of information which pattern the objects in U belong to makes it impossible to calculate this quantity directly. We are forced to introduce additional assumptions:

(1) given an object z of the mixed sample, the nearest object in U belongs to the same pattern as z;

(2) the nearest rival object in U lies at the distance at most r^*;

(3) as the distance from an object to an pattern we take the distance from the object to the nearest representative of this pattern.

The technique for calculating the function of rival similarity from a mixed sample under these assumptions is described in [12]. This technique underlies the FRiS-TDR algorithm [13] which constructs a system of stolps for a mixed sample. The share d of classified objects in the sample can take an arbitrary value between 0 and 1. For $d = 0$, we in fact solve the taxonomy problem; for $d = 1$, the problem of constructing decision rules; and, in the intermediate case, the semi-supervised learning problem. Thus, we have developed a unified approach to these three problems on the basis of the FRiS-function.

6. INCREASE OF THE COMPACTNESS OF DATA BY CENSORING OBJECTS. THE FRiS-CENSOR ALGORITHM

We can simplify the description and increase compactness by refining the data, i.e., removing the "untypical" objects which distort the representation of the sample and affect the choice of standards. The significant differences in the properties of these objects from the properties of the remaining objects of the pattern can be explained by their uniqueness, but more often the reason lies in the influence of neglected factors like sensor failures, errors of data entry into the protocol, and others. Sometimes there are objects which are not "mistaken" but lie on the periphery of the distribution and turn out deep in the intersection zone with the neighboring patterns. They also can unjustifiably complicate the decision rules.

In order to censor the outliers, we can apply the FRiS-Censor algorithm [7] which includes the FRiS-Stolp algorithm as its part and uses the measure of FRiS-compactness of the patterns as the criterion controlling the process of increasing the compactness of data.

Suppose that we are given some two patterns A and B expressed as tuples of M_A and M_B objects, and put $M = M_A + M_B$. Estimate the compactness of A and B using (2) and (3). Let M^* denote the number of objects in the training sample remaining after the current sample reduction step. We will use $(M^*/M)^\alpha$, with $\alpha \geq 0$, as the penalty for excluding objects from the training sample. Taking this into account, estimate the FRiS-compactness $H_{A|B}$ of the patterns at each sample reduction step as

$$H_{A|B} = (M^*/M)^\alpha \sqrt{C_{A|B} C_{B|A}}.$$

We choose the optimal value of α using the method of computer modelling by comparing the results of application of the FRiS-Censor algorithm for different α values. Let $d \in [0, 1]$ be the maximal share of objects of the training sample which we can exclude, and m^*, the maximal number of objects in a removed cluster.

We tested the FRiS-Censor algorithm on a model recognition problem of the two patterns each of which amounted to the superposition of several (from 2 to 4) normally distributed clusters in a two-dimensional space of features. We considered 10 distributions with different dispersions of clusters, coordinates of expected values, and numbers of objects in the clusters, which influenced the FRiS-compactness of the patterns. Every pattern was represented by 250 objects. For every distribution, we 100 times randomly divided the sample into two parts: the training part (50 objects of the first pattern and 50 objects of the second pattern) and the control part (200 objects of each pattern). The total number of experiments for various numerical values of the initial data reached 1000.

For all data, we ran the algorithm using the parameters: $\alpha = 0, 1, \ldots, 9$, $m^* = 4$, and $d = 0.15$; thus, out of 100 objects of the training sample at most 15 objects could be removed.

At every stage of the training sample reduction, we used the FRiS-Stolp algorithm to construct its description as a collection of stolps. This enabled us to calculate the FRiS-compactness of the patterns. The step of the algorithm on which this quantity was maximal was regarded as the best one. Using the stolps chosen at this step, we constructed a decision rule and applied it to recognize 400 objects not included into training. The reliability P of this recognition (in percent), averaged over all 1000 experiments, enabled us to determine the optimal value $\alpha = 5$.

The experiments show that the increase of the FRiS-compactness of the training sample by censoring objects in more than 99(%) of cases leads to a higher quality of recognition. The refined sample is described by a simpler decision rule, which increases the reliability of recognition of the control sample. Fig. 4 depicts the distributions of the reliability P (%) of the recognition of the control sample. The y-axis corresponds to the absolute number of experiments N (out of 1000) in which this reliability P was achieved. Curve 1 corresponds to reliability without a prior increase of the FRiS-compactness, and the average value is equal to 91.6(%), while curve 2, to the reliability with the use of the increase of FRiS-compactness. Here the average value is equal to 95.9(%). The average value d^* for which the maximal value of the criterion $H_{A|B}$ was attained equals 12.7(%).

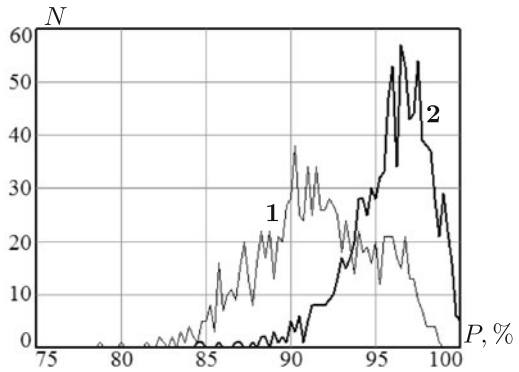

Fig. 4. The distribution of the reliability P of recognition of the control sample

7. EXCLUSION OF REDUNDANT FEATURES. THE FRiS-GRAD ALGORITHM

In the presently dominating problems, the number N of features exceeds the number M of objects by orders of magnitude. Moreover, the information useful for solving a certain classification problem is usually present in several features $n \ll N$. The choice of these n features not only enables us subsequently to substantially reduce the expense of computer resources, but also increases the compactness of the patterns and the reliability of their recognition. The features can depend on each other, which makes it impossible to choose from the estimates of individual informativity of each feature a subset as a list of n most relevant features. If n is given then the exact solution would require checking all combinations of n out of N features, which is impossible in real problems. For this reason, some heuristic algorithms of directed search are used.

The GRAD algorithm [14], which we developed, uses the trick proposed in [15]. To begin with, the brute-force search forms a system of relevant features, called granules, of small dimension. We use these granules as input generalized features for the AdDel algorithm [10], which amounts to a combination of the two available greedy algorithms: Addition [16] and Deletion [17]. These algorithms yield the optimal solution at each step without guaranteeing the global optimum.

The AdDel algorithm in its forward step (the Addition algorithm) gathers a certain number of relevant features (granules of features), and then excludes a part of them in an backward pass (the Deletion algorithm). The Addition and Deletion algorithms keep alternating until a specified number n of features (granules of features) is reached. In the FRiS-GRAD algorithm, which uses the granules instead of the separate features, some features in the resulting system can occur more than once.

Our experiments show that, as the number of features increases, the quality of recognition firstly increases, then the growth stops, and then it starts to decrease on account of the addition of redundant, noisy features. The inflection point of the quality curve enables us to automatically determine the number of relevant features in a system.

We can estimate the informativity of a feature or a system of features using different methods. The quality of solution to this problem depends on how universal and suitable for the problem a criterion we use. In the FRiS-GRAD algorithm [18] for the feature selection, we use the FRiS-compactness as the informativity criterion. This criterion applies for every distribution and every ratio of M and N.

During the calculation of the FRiS-compactness we simultaneously choose a system of stolps. Therefore, we can interpret the FRiS-GRAD algorithm as a data compression algorithm on accounting the decrease in the number of objects in the sample and the number of features describing the sample. We can then use this reduced description of the sample as a set of stolps in the space of relevant features to solve the recognition problem. An object z belongs to the pattern the similarity to whose stolp in the space of the chosen relevant features turns out the greatest, while we regard the similarity as the probability that the decision made is correct.

Table 1. The results of solving nine problems

Problem	N	M_1/M_2	Record results out of 40	Results FRiS–GRAD	Rating FRiS-GRAD
ALL1	12625	95/33	100.0	100.0	1
Leuk	7129	47/25	95.85	100.0	1
Prost	12625	50/53	90.19	93.13	1
DLBCL	7129	58/19	94.30	93.51	3
Colon	2000	22/40	88.60	89.52	1
ALL4	12625	26/67	82.06	83.87	1
Myel	12625	36/137	82.90	81.45	2
ALL3	12625	65/35	59.58	73.82	1
ALL2	12625	24/91	78.23	80.75	1
Middle			85.75	88.45	

8. EXAMPLES OF THE INCREASE OF COMPACTNESS OF ILL-CONDITIONED DATA

In order to estimate the efficiency of the FRiS-GRAD algorithm, we ran a large-scale test on nine medical problems. The objects were patients with various diseases and the features were gene expressions obtained using biochips. A specific of these problems is that they are ill-conditioned: the number of features exceeds by several orders of magnitude the number of objects in the sample.

We compared the results of the work of the algorithm to the previous results obtained by the four most frequently used recognition algorithms (support vector machines, between-group analysis, Bayes classifier, and k-nearest-neighbors) in the relevant subspaces chosen by 10 available feature selection algorithms.

We estimated the quality of the algorithms applying cross-validation: we used 50(%) of the sample for training, and the remaining 50(%) to estimate the reliability of recognition. We took all results except those related to FRiS-GRAD from [19], which for every problem presents 40 distinct versions of solutions obtained by all possible combinations of the algorithms. For the comparison, we chose the best results in each problem.

Table 1 presents the results of our comparison. Here we show the names of the problems, the dimensions N of the feature spaces, the ratios of the number M_1 of objects of the first pattern to the number M_2 of objects of the second pattern, and three columns of results. The last column shows the places taken by the results of solving all nine problems by the FRiS-GRAD algorithm.

The result obtained by each of 10 feature selection methods indicates its rating: the best result receives rating 1, and the worst receives 10. Summing the ratings obtained by a method over all problems, we find its overall rating. Table 2 presents the results of this calculation, showing in its last rows the sum of the ratings of the places taken by FRiS-GRAD. We performed the same analysis using four decision rules. Table 3 presents its results.

Our comparison shows the high efficiency of the FRiS-GRAD algorithm against the background of the available analogs. The compressed description of the sample constructed by this algorithm yields better results of recognition than the results achieved by other algorithms on the entire sample.

Table 2. The sums of ratings of feature selection methods

Feature selection method	Rating
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
maxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	12

Table 3. The sum of ratings of decision rules

Decision rule	Rating
Between group analysis (BGA)	35
K-nearest-neighbors (kNN)	32
Naive Bayes classification (NBC)	25
Support vector machines (SVM)	19
FRiS-Stolp	12

CONCLUSION

The passage from a binary relation to a ternary relation for describing a measure of rival similarity of objects enables us to introduce substantial modifications into the methods of cognitive data mining.

The use of the FRiS-function enables us to simplify and improve the existing data mining algorithms, increase their noise immunity, make them insensitive to the shape of the distribution of the patterns and to the ratio of the numbers of objects and features.

We have managed to state and solve new data mining problems: to obtain a quantitative estimate for compactness, to censor the training sample, to obtain a universal classification of objects with an automatic choice of the number of clusters, and to select features using a new informativity criterion with the automatic determination of the number of features in the system.

ACKNOWLEDGMENTS

The authors were supported by the Russian Foundation for Basic Research (project no. 11–01–00156).

REFERENCES

1. A. G. Arkad'ev and E. M. Braverman, *Machine Learning Object Classification* (Nauka, Moscow, 1971) [in Russian].
2. K. V. Vorontsov and A. O. Koloskov, "Profiles of Compactness and Isolation of Reference Objects in Metric Classification Algorithms," Art. Intelligence No. 2, 30–33 (2006).
3. M. I. Schlesinger, "On the Spontaneous Sharing of Images," in *Reading Machines and Pattern Recognition* (Naukova Dumka, Kiev, 1965), pp. 46–61.
4. N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko, "Methods of Recognition Based on the Function of Rival Similarity," Pattern Recognition and Pattern Analysis **18** (1), 1–6 (2008).
5. K. Kira and L. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," in *Proceedings of the 10th Conference on Artificial Intelligence (AAAI-92)* (AAAI Press, Menlo Park, 1992), pp. 129–134.
6. P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," J. Comput. Appl. Math. **20**, 53–65 (1987).
7. N. G. Zagoruiko and O. A. Kutnenko, "A Quantitative Measure of Compactness of Images and Method of Its Increase," in *Proceedings of 9 International Conference "Intellectualization of Information Processing," Budva, September 16–22, 2012* (Torus Press, Moscow, 2012), pp. 29–32.
8. I. A. Borisova, V. V. Dyubanov, N. G. Zagoruiko, and O. A. Kutnenko, "Similarities and Compactness," in *Proceedings of 14 All-Russia Conference "Mathematical Methods for Pattern Recognition"* (Moscow, 2009), pp. 89–92.
9. I. A. Borisova, "The Algorithm Taxonomy FRiS-Tax," Scientific Bulletin of Novosibirsk State Technical Univ. No. 3, 3–12 (2007).
10. N. G. Zagoruiko, *Advanced Methods of Data and Knowledge Analysis* (Inst. Mat., Novosibirsk, 1999) [in Russian].
11. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of 5 Berkley Symposium on Mathematical Statistics and Probability*, Vol. 1 (Univ. Press, California, 1967), pp. 281–297.
12. I. A. Borisova, 'Calculation of FRiS-Function over Mixed Dataset in the Task of Generalized Classification," in *Proceedings of 3 International Conference on Inductive Modeling (ICIM'2010)* (Yevpatoria, 2010), pp. 44–50.
13. I. A. Borisova and N. G. Zagoruiko, "FRiS-TDR Algorithm for Solving the Generalized Problem of Taxonomy and Identification," in *Proceedings of All-Russia Conference UMBRELLA-2009*, Vol. 1 (Novosibirsk, 2009), pp. 93–102.
14. N. G. Zagoruiko and O. A. Kutnenko, "GRAD Algorithm for Feature Selection," in *Proceedings of 8 International Conference "The Use of Multivariate Statistical Analysis in Economics and Assessing the Quality"* (Moscow State Univ. Econom. Statist. Inform., Moscow, 2006), pp. 81–89.
15. P. Pudil, J. Novicova, and J. Kittler, "Floating Search Methods in Feature Selection," Pattern Recognition Lett. (1994) **15** (11), 1119–1125.
16. T. Merill and O. M. Green, "On the Effectiveness of Receptions in Recognition Systems," IEEE Trans. Inform. Theory. **IT-9**, 11–17 (1963).
17. *Automatic Pattern Recognition* (Kiev Higher Artillery Engineering School, Kiev, 1963) [in Russian].
18. N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko, "Attribute Selection Through Decision Rule Construction (Algorithm FRiS-GRAD)," in *Proceedings of 9 International Conference on Pattern Recognition and Pattern Analysis: New Information Technologies*, Vol. 2 (Nizhnii Novgorod, 2008), pp. 335–338.
19. I. Jeffery, D. Higgins, and A. Culhane, "Comparison and Evaluation of Methods for Generating Differentially Expressed Gene Lists from Microarray Data," BMC Bioinformatics **7**, 359–374 (2006).