Grouping of Objects in a Space of Heterogeneous Variables with the Use of Taxonomic Decision Trees

V. B. Berikov

Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences,
pr. akademika Koptyuga 4, Novosibirsk, 630090 Russia
e-mail: berikov@math.nsc.ru

Abstract—A problem of classification of objects in the presence of heterogeneous (qualitative, ordinal, nominal, and Boolean) variables is considered. Taxonomic decision trees are used to solve the problem. A quality criterion for a tree is introduced that is based on the Bayesian estimate of the Kullback–Leibler distance between distributions. Statistical modeling is applied to show the efficiency of an algorithm for constructing a tree that uses this criterion.

Keywords: automatic classification, heterogeneous variables, decision tree.
DOI: 10.1134/S1054661811040043

INTRODUCTION

The problem of cluster analysis (taxonomy, grouping of objects according to the similarity of their characteristics, unsupervised classification) can be formulated as follows. Suppose given a set of objects described by a collection of certain variables. From these objects, it is required to form a relatively small number of clusters (taxons, groups, or classes) so that the quality criterion for grouping takes the best value. By the quality criterion is usually meant a certain functional depending on the dispersion of objects in a group and on the distances between groups.

Often one faces the need to cluster objects described by heterogeneous variables, i.e., variables measured on different scales: interval, ordinal, nominal, and Boolean scale. An example of such a problem is given by the analysis of medical data related to patients characterized by both quantitative (age, weight, blood cholesterol level, etc.) and qualitative (sex, profession, attitude to smoking, etc.) features.

One can list the following main methods for solving the problems of cluster analysis in the case of heterogeneous variables. There are methods based on introducing the distance between objects in a heterogeneous feature space. For example, in [1] the authors proposed a k-prototype algorithm in which a combination of Euclidean and Hamming metrics with some weights are used to calculate distances. By the quality criterion for grouping is meant the total dispersion of objects with respect to the "centers" of groups ("prototypes"). However, when introducing a metric in a heterogeneous space, one faces complicated methodical questions, and the definition of the best weights of the variables still remains an unsolved problem.

Another method consists in reducing the analysis of heterogeneous variables to the analysis of variables of the same type. For instance, in [2] the authors proposed a cluster analysis algorithm based on evaluating the parameters of a mixture of polynomial distributions defined on combinations of discretized variables (i.e., the range of quantitative variables is preliminarily partitioned into intervals of fixed length). A disadvantage of such a method is the loss of information on the closeness of objects, as well as the fact that, in the case of high dimension of the space and a large number of intervals (names), one faces a serious problem of reliability of estimating by a limited number of observations.

Many authors (see, for example, [3]) describe a procedure based on an ensemble of algorithms of cluster analysis each of which carries out grouping in a subspace of variables of the same type. Approaches to the construction of an ensemble of clustering algorithms are described in [4]. In spite of a large number of experimental verifications of the advantage of the ensembles of clustering algorithms, a sufficiently full theoretical substantiation of their efficiency is still missing. Some questions related to the theoretical analysis of the quality of ensembles of clustering algorithms (by a pairwise classification model) were considered in [5].

One of possible approaches to the cluster analysis in the presence of heterogeneous variables is based on the application of decision trees. Decision trees are often used in classification and prediction problems in the case of heterogeneous variables; they allow one to obtain an easily interpretable logical model of grouping and to select the most informative factors and do not require specifying a metric in a heterogeneous

Received August 29, 2010

space. Decision trees were first used in the cluster analysis of heterogeneous data in [6]. Various modifications of algorithms for constructing decision trees for a problem of cluster analysis are described in [5, 7, 8] and other publications.

A specific feature of the approach based on decision trees is that it allows one not only to obtain a decomposition of a given set of objects into clusters but also to construct a hierarchic tree that describes the structure of the decomposition and allows one to refer an arbitrary new object to the taxons obtained (or to point out that this object is not typical and, possibly, belongs to an unknown class or represents noise, or background).

When constructing a taxonomic decision tree, one performs a directed search for the best variant by a given quality criterion. To this end, one can apply various modifications of an LRP-type greedy algorithm, a recursive R algorithm [9], etc. These algorithms involve a quality criterion based on the concept of a relative volume of a taxon. However, a disadvantage of this criterion is that it is insensitive to the number of objects that make up a cluster. Hence, other conditions being equal, one may prefer nonrepresentative clusters that consist of a small number of objects. The aim of the present study is to develop a new quality criterion that allows one to take into account the number of objects in groups. To this end, we propose the application of a combination of information and Bayesian approaches.

The paper is organized as follows. In the first section, we give the main concepts used in the paper. In the second section, we introduce information quality criteria for a taxonomic decision tree that are based on the Kullback–Leibler distance between distributions. We consider both a frequency estimate for this distance and an estimate obtained on the basis of a Bayesian approach. To obtain the Bayesian estimate, we find an expression for the expected entropy of an a posteriori distribution of classes. In the third section, we describe the method and the results of experimental investigation of algorithms based on the criteria introduced. In the Conclusions, we summarize the main results of the paper.

1. THE BASIC CONCEPTS

Suppose given a set $s = \{o^{(1)}, ..., o^{(N)}\}$ of objects chosen from the statistical population. Each object is described by an ensemble of variables $X_1, ..., X_m$. This ensemble may include variables of different types (quantitative and qualitative, by which we mean nominal and Boolean, as well as ordinal, variables). Let D_j stand for the set of values of the variable X_j (which is an interval of the real axis in the case of a quantitative variable, or a finite set of values (names) in the case of a qualitative variable). Let $D = \prod_j D_j$. Denote by $x = x(o) = (x_1(o), ..., x_m(o))$ an ensemble of observations of the variables for an object o, where $x_j(o)$ is the value of the variable X_j for the given object. We will represent the ensemble of observations corresponding to the set of objects as a data table with N rows and m columns.

In a cluster analysis problem, one has to partition objects into a certain number K ($K \ll N$) of clusters so that a given quality criterion for groping takes an optimal value. The number of classes may be either chosen in advance or not specified (in the latter case, the optimal number of clusters should be defined automatically).

In some problems, one should not only partition objects into similar groups but also obtain a rule that would allow one to refer an arbitrary new object to a certain class with number Y. To this end, one uses the so-called taxonomic decision function, by which we mean a mapping $D \longrightarrow D_Y$, where $D_Y = \{0, 1, ..., K\}$ is a set of numbers of classes such that $Y = 0$ implies that an object does not belong to any of the clusters found. Such objects will be said to be atypical (or noise objects).

Note that the concept of a taxonomic decision function was introduced in [10], where it was used for solving a pattern recognition problem.

The main problem is as follows. It is required to construct a taxonomic decision function that belongs to a given class and is optimal by a certain criterion. To solve this problem, one should define an appropriate class of decision functions, define a quality criterion, and formulate an algorithm for choosing an optimal function. Below we consider the above-mentioned questions as applied to the case of taxonomic decision trees.

1.1. A Taxonomic Decision Tree

Consider a tree in which each internal vertex (node) is assigned a variable X_j and the branches emanating from this vertex correspond to a statement of the form $X_j(o) \in E_j^{(i)}$, where o is an object; $i = 1, 2, ..., \nu$, $\nu \geq 2$ is the number of branches emanating from the given vertex; and $E_j^{(1)}, ..., E_j^{(\nu)}$ are pairwise disjoint subsets of the set D_j (intervals of values in the case of a quantitative variable). Each kth leaf (terminal node) of the tree corresponds to a group of objects of sample $C^{(k)} = \{o^{(i_1)}, ..., o^{(i_{n^{(k)}})}\}$, where $n^{(k)}$ is the number of objects in the group and $k = 1, ..., K$. The objects of a given group satisfy a chain of statements that are verified along a path from the root vertex to this leaf, i.e., a logical assertion of the form

$$\text{If } X_{j_1}(o) \in E_{j_1}^{(i_1)} \text{ AND}$$

$$X_{j_2}(o) \in E_{j_2}^{(i_2)} \text{ AND } ... \text{ AND } X_{j_{q_k}}(o) \in E_{j_{q_k}}^{(i_{q_k})},$$

$$\text{then the object } o \text{ belongs to the } k\text{th class,}$$

where q_k is the length of the given chain and $o \in C^{(k)}$.

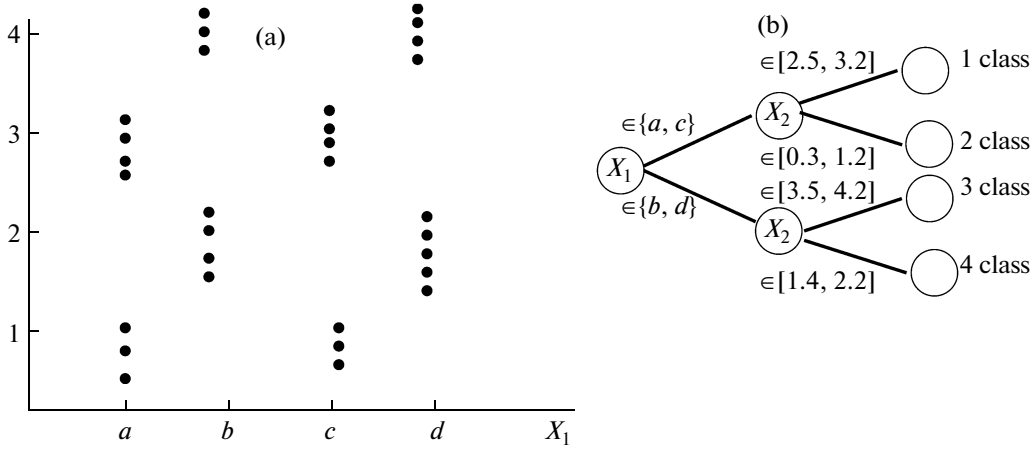

Fig. 1. Examples of (a) a clusterization problem and (b) a taxonomic decision tree.

By a taxon $T^{(k)}$ corresponding to the kth leaf of a tree we mean a rectangular subdomain containing objects from the group $C^{(k)}$ in a multidimensional space:

$$T^{(k)} = T(C^{(k)}) = T_1^{(k)} \times \ldots \times T_j^{(k)} \times \ldots \times T_m^{(k)},$$

where

$$T_j^{(k)} = \{X_j(o) | o \in C^{(k)}\}$$

for the qualitative variable X_j and

$$T_j^{(k)} = [\min_{o \in C^{(k)}} X_j(o); \max_{o \in C^{(k)}} X_j(o)]$$

for a quantitative variable X_j.

For a new observation x, by verifying the statements corresponding to the tree, we seek a taxon $T^{(k)}$ that contains this observation. To x, we assign the kth class. If x is not contained in any of the taxons available, then we assign it the value $Y = 0$.

The tree described is called a *taxonomic decision tree*. The vertices of the tree correspond to a certain nested hierarchy of subsets of objects. Note that the tree does not necessarily contain the entire original ensemble of variables.

An illustrative example of the disposition of observations in a space of two variables is shown in Fig. 1a. Here X_1 is a nominal variable, while X_2 is a quantitative variable. An example of a taxonomic decision tree that partitions objects into $K = 4$ groups is shown in Fig. 1b.

1.2. Quality Criterion for a Decision Tree for a Heterogeneous Space

Suppose that a partition $G = \{C^{(1)}, ..., C^{(k)}, ..., C^{(K)}\}$ of a set of objects into groups is formed in accordance with a taxonomic decision tree. In the case of quantitative variables, by a quality criterion for grouping is usually meant the total dispersion of points with respect to the centers of clusters. However, in the case of a heterogeneous space of variables, such a criterion is inapplicable.

In [6], for a heterogeneous space, the authors proposed a quality criterion for grouping that is based on the minimization of the total relative volume of taxons

$$Q(G) = \sum_{k=1}^{K} V^{(k)},$$

where $V^{(k)} = V(T^{(k)}) = \prod_{j=1}^{m} \frac{|T_j^{(k)}|}{|D_j|}$ is the relative volume of the taxon $T^{(k)}$; $|\cdot|$ stands for the cardinality of the corresponding set or the length of the interval.

One can easily verify that the minimization of this criterion is equivalent to the maximization of the criterion

$$Q'(G) = \sum_{k=1}^{K} (\left|\hat{P}(T^{(k)}) - P_u(T^{(k)})\right|),$$

where $\hat{P}(T^{(k)}) = \frac{n^{(k)}}{N}$ is the relative frequency of falling into the taxon $T^{(k)}$ and $P_u(T^{(k)}) = V(T^{(k)})$ is the probability of falling into the taxon provided that we deal with independent random variables each of which has a uniform distribution. Note that the use of the absolute value in the formula for $Q'(G)$ is not necessary if we assume that we consider only taxons the density of points in which is higher than the mean density over the entire space (i.e., $\frac{n^{(k)}}{|T^{(k)}|} > \frac{N}{|D|}$, where $|T^{(k)}| = \prod_j |T_j^{(k)}|$ and $|D| = \prod_j |D_j|$).

Thus, when using this criterion, one seeks for groups whose distribution most strongly differs from the uniform distribution in the above sense.

In [9], the authors used a modification of the criterion (a regularizing criterion) in which the necessary number of clusters is not defined and one seeks for a certain compromise between the total volume and the number of groups:

$$Q_R(G) = \sum_{k=1}^{K} V^{(k)} + \alpha\frac{K}{N},$$

where α is an heuristic parameter. The experimental investigations carried out in [9] have shown that there exists a range of values of this parameter (usually, α is defined on the interval from one to two) in which the solutions have acceptable quality.

1.3. An Algorithm for Constructing a Tree

It is well known that the problem of constructing an optimal decision tree is NP hard in the general case. Therefore, as a rule, one applies an approximate algorithm for searching for an optimal tree in which a directed search for variants is applied.

In [9], the authors describe a successive branching algorithm LRP and a recursive algorithm (the R method). At each step of the algorithm LRP, a certain group of objects that corresponds to a vertex of a tree is partitioned into two new subgroups. The partition is performed with the use of the quality criterion for grouping Q; i.e., the total volume of the taxons obtained is minimized. A vertex for which the relative volume of the corresponding taxon is greater than a given parameter is considered to be promising for further branching. The branching is continued either until there remain no more promising vertices or until a given number of groups is obtained.

The R method involves a recursive scheme of searching for different variants of a tree to a given depth R. First, in each of the levels of the tree, one constructs a maximum possible number of vertices (defined by a sample). Then these vertices are successively united until an optimal value of the quality criterion is reached. In this case, one applies a regularizing criterion Q_R. The tree obtained is not necessarily binary.

By increasing the parameter R, one can increase the depth of the search for variants, which allows one to take into account more complicated dependence between the variables (in this case, the operation time and the required memory volume increase). A distinctive feature of the algorithm is that the number of branches emanating from each vertex is not fixed in advance, and one seeks for the optimal number of such branches. The details of the algorithm are described in [9].

2. THE PROPOSED QUALITY CRITERION

The quality criteria for grouping Q and Q_R considered above do not take into account the composition of groups (i.e., the number of objects included in a group). Let us introduce another criterion, which is based on the Kullback–Leibler distance between the distribution of the probability to fall into taxons and a uniform distribution. Define an empty (or noise) domain $T^{(0)} = D\backslash\{T^{(1)} \cup \ldots \cup T^{(k)}\}$. The Kullback–Leibler distance is defined as

$$\rho_{KL} = \sum_{k=0}^{K} P(T^{(k)}) \ln\frac{P(T^{(k)})}{P_u(T^{(k)})},$$

where $P(T^{(k)})$ is the probability that a randomly chosen observation belongs to the domain $T^{(k)}$, $k = 0, 1, \ldots, K$, and the domain $T^{(0)}$ satisfies the equality $P_u(T^{(0)}) = 1 - \sum_{k=1}^{K} V^{(k)}$. In addition, assume that $0 \cdot \ln 0 = 0$. To evaluate the probability to fall into subdomains, we can use appropriate frequencies. Denote a criterion based on the frequency estimate for the above distance by $Q_{\mathrm{FKL}}(G)$.

It is well known that frequency estimates have greater error for a problem of higher dimension and a relatively small number of objects. To increase the accuracy of estimates, one can additionally invoke a priori knowledge available for the researcher. We will use the earlier developed Bayesian model of classification by a finite set of events [9] in which an a priori distribution is defined on the set of states of nature according to expert information.

Consider a discrete random variable X with a set of unordered values $D_X = \{u^{(0)}, u^{(1)}, \ldots, u^{(K)}\}$, where $u^{(k)}$ is the kth value (cell) corresponding to the subdomain $T^{(k)}$, $k = 0, 1, \ldots, K$. For convenience, we encode the variable X by the numbers of cells. Let $p^{(k)}$ be the probability of the event "$X = k$," such that $p^{(k)} \geq 0$, $k = 0, 1, \ldots, K$, and $\sum_{k=0}^{K} p^{(k)} = 1$. Let $n^{(k)}$ be the number of observations corresponding to the kth cell; $\sum_{k=0}^{K} n^{(k)} = N$ (note that $n^{(0)} = 0$ in the absence of noise). Denote an observed vector of frequencies by $s = (n^{(0)}, n^{(1)}, \ldots, n^{(K)})$. Let S stand for a random vector of frequencies that obeys a polynomial distribution with the parameter vector $\theta = (p^{(0)}, p^{(1)}, \ldots, p^{(K)})$. Consider a family of polynomial distribution models defined by a set of parameters $\Lambda = \{\theta\}$. This family (class of distributions) is also called a set of models of the states of nature.

We apply a Bayesian approach: *Suppose that a random variable* $\Theta = (P^{(0)}, P^{(1)}, ..., P^{(K)})$ *with a known a priori distribution* $p(\theta)$ with $\theta \in \Lambda$ *is defined on* Λ. We will assume that Θ satisfies the Dirichlet distribution, $\Theta \sim \mathrm{Dir}(\mathbf{d})$: $p(\theta) = \frac{1}{Z}\prod_{k=0}^{K}(p^{(k)})^{d^{(k)}-1}$, where $\mathbf{d} = \{d^{(0)}, d^{(1)}, ..., d^{(K)}\}$ and $d^{(k)} > 0$ are some real numbers that express expert knowledge on the distribution Θ ($k = 0, 1, ..., K$) and Z is a normalization constant ($Z = \frac{\prod_{k=0}^{K}\Gamma(d^{(k)})}{\Gamma(D)}$, where $\Gamma(\cdot)$ is a gamma function and $D = \sum_{k=0}^{K} d^{(k)}$). The more convinced an expert is that the frequency of the kth taxon should be relatively high, the greater the value of $d^{(k)}$. In the absence of knowledge about a priori preferences, one can apply a uniform a priori distribution ($\mathbf{d} = 1$).

Consider the entropy of a distribution as a function of θ:

$$H(\theta) = -\sum_{k=0}^{K} p^{(k)} \ln p^{(k)}.$$

We will call the mathematical expectation of the entropy $\overline{H} = \mathbf{E}_{\Theta} H(\Theta)$, where the averaging is performed over the set Λ, the expected entropy of a priori distribution.

Proposition 1. *Suppose that the above assumptions hold. Then the expected entropy is given by*

$$\overline{H} = \psi(D+1) - \sum_{k=0}^{K} \frac{d^{(k)}}{D}\psi(d^{(k)}+1),$$

where $\psi(z) = \frac{d}{dz}\ln\Gamma(z)$ *is the digamma function.*

Proof. We have

$$\mathbf{E}_{\Theta} H(\Theta) = -\frac{1}{Z}\int_{\Lambda}\sum_{k=0}^{K} p^{(k)} \ln p^{(k)} p(\theta) d\theta$$

$$= -\frac{1}{Z}\sum_{k=0}^{K}\int_{0}^{1}(p^{(k)})^{d^{(k)}} \ln p^{(k)} \int_{\Lambda^{(k)}}\prod_{l\in I^{(k)}}(p^{(l)})^{d^{(l)}-1}\prod_{l\in I^{(k)}} dp^{(l)} dp^{(k)},$$

where $I^{(k)} = \{l | l = 0, ..., K, l \neq k\}$ and $\Lambda^{(k)} = \left\{p^{(l)} \Big| \sum_{l\in I^{(k)}} p^{(l)} = 1 - p^{(k)}\right\}$, $k = 0, ..., K$.

Let us apply an integral formula that follows from the generalized Liouville formula [11, p. 397]:

$$\int_{\substack{x_1,...,x_{n-1}:\\ x_i\geq 0\\ \sum_{i=1}^{n-1} x_i \leq h}} \prod_{i=1}^{n-1} x_i^{d_i-1}\left(h - \sum_i x_i\right)^{d_n-1} dx_1 \ldots dx_{n-1}$$

$$= \frac{\prod_{i=1}^{n}\Gamma(d_i)}{\Gamma\left(\sum_{i=1}^{n} d_i\right)} h^{\sum_{i=1}^{n} d_i - 1},$$

where $d_1, ..., d_n$ are positive real numbers. We obtain

$$\mathbf{E}_{\Theta} H(\Theta) = -\frac{1}{Z}\sum_{k=0}^{K}\int_{0}^{1}(p^{(k)})^{d^{(k)}}$$

$$\times \ln p^{(k)} \frac{\prod_{l\in I^{(k)}}\Gamma(d^{(l)})}{\Gamma\left(\sum_{l\in I^{(k)}} d^{(l)}\right)}(1-p^{(k)})^{\sum_{l\in I^{(k)}} d^{(l)}-1} dp^{(k)}$$

$$= -\frac{1}{Z}\sum_{k=0}^{K}\frac{\prod_{l\in I^{(k)}}\Gamma(d^{(l)})}{\Gamma(D-d^{(k)})}$$

$$\times \int_{0}^{1}(p^{(k)})^{d^{(k)}}(1-p^{(k)})^{D-d^{(k)}-1}\ln p^{(k)} dp^{(k)}.$$

Now, we apply the following integral formula ([12, p. 552]):

$$\int_{0}^{1} x^{\mu-1}(1-x)^{\nu-1}\ln x dx = \mathrm{B}(\mu,\nu)[\psi(\mu) - \psi(\mu+\nu)],$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function and $\mu, \nu > 0$. We obtain

$$\mathbf{E}_{\Theta} H(\Theta) = -\frac{1}{Z}\sum_{k=0}^{K}\frac{\prod_{l\in I^{(k)}}\Gamma(d^{(l)})}{\Gamma(D-d^{(k)})}$$

$$\times \mathrm{B}(d^{(k)}+1, D-d^{(k)})[\psi(d^{(k)}+1)$$

$$-\psi(d^{(k)}+1+D-d^{(k)})].$$

Using the property $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, after transformations we obtain

$$\mathbf{E}_\Theta H(\Theta) = -\frac{\Gamma(D)}{\Gamma(D+1)} \times \sum_{k=0}^{K} \frac{\Gamma(d^{(k)}+1)}{\Gamma(d^{(k)})}[\psi(d^{(k)}+1) - \psi(D+1)]$$
$$= \psi(D+1) - \sum_{k=0}^{K} \frac{d^{(k)}}{D}\psi(d^{(k)}+1),$$

which was to be proved. Note that $\psi(z) \approx \ln z$ holds for large z.

Suppose that a sample (an observable vector of frequencies) s is obtained. By a property of the Dirichlet distribution, the a posteriori distribution is given by $\Theta|_S \sim \mathrm{Dir}(d^{(0)} + n^{(0)}, d^{(1)} + n^{(1)}, \ldots, d^{(K)} + n^{(K)})$.

We will call the mathematical expectation of the entropy $\overline{H}_s = \mathbf{E}_{\Theta|s} H(\Theta)$, where the averaging is performed over all models of the state of nature according to an a posteriori distribution, the expected entropy of the a posteriori distribution. Proposition 1 implies the following proposition.

Proposition 2. *The expected entropy of a posteriori distribution is given by*

$$\overline{H}_s = \psi(D+N+1) - \sum_{k=0}^{K} q^{(k)}\psi(d^{(k)} + n^{(k)} + 1),$$

where $q^{(k)} = \frac{d^{(k)} + n^{(k)}}{D+N}$, $k = 0, 1, \ldots, K$.

The quantities $q^{(k)}$ can be interpreted as Bayesian estimates for the probability to fall into the subdomains $T^{(0)}, \ldots, T^{(K)}$.

Consider the Kullback–Leibler distance as a function of a state θ of nature:

$$\rho_{\mathrm{KL}}(\theta) = \sum_{k=0}^{K} p^{(k)} \ln \frac{p^{(k)}}{V^{(k)}}.$$

Proposition 3. *The a posteriori mathematical expectation* $E_{\Theta|s}\rho_{\mathrm{KL}}(\Theta|s)$ *is defined as follows*:

$$\mathbf{E}_{\Theta|s}\rho_{\mathrm{KL}}(\Theta|s) = -\left(\overline{H}_s + \sum_{k=0}^{K} q^{(k)} \ln V^{(k)}\right).$$

Proof. We have

$$\mathbf{E}_{\Theta|s}\rho_{\mathrm{KL}}(\Theta|s) = \mathbf{E}_{\Theta|s} \sum_{k=0}^{K} p^{(k)} \ln \frac{p^{(k)}}{V^{(k)}}$$
$$= \mathbf{E}_{\Theta|s} \sum_{k=0}^{K} p^{(k)} \ln p^{(k)} - \mathbf{E}_{\Theta|s} \sum_{k=0}^{K} p^{(k)} \ln V^{(k)}.$$

Using Proposition 2, we obtain

$$\mathbf{E}_{\Theta|s}\rho_{\mathrm{KL}}(\Theta|s) = -\overline{H}_s - \sum_{k=0}^{K} \ln V^{(k)} \mathbf{E}_{\Theta|s} p^{(k)}$$
$$= -\overline{H}_s - \sum_{k=0}^{K} \ln V^{(k)} \frac{\Gamma(D+N)}{\prod_{l \in I^{(k)}} \Gamma(d^{(l)} + n^{(l)})}$$
$$\times \int_\Lambda p^{(k)} \prod_{l \in I^{(k)}} (p^{(l)})^{d^{(l)} + n^{(l)} - 1} dp^{(0)} \ldots dp^{(K)}$$
$$= -\overline{H}_s - \sum_{k=0}^{K} \ln V^{(k)} \frac{\Gamma(D+N)}{\prod_{l \in I^{(k)}} \Gamma(d^{(l)} + n^{(l)})}$$
$$\times \frac{\Gamma(d^{(k)} + n^{(k)} + 1) \prod_{l \in I^{(k)}} \Gamma(d^{(l)} + n^{(l)})}{\Gamma(D+N+1)}$$
$$= -\overline{H}_s - \sum_{k=0}^{K} \ln V^{(k)} \frac{d^{(k)} + n^{(k)}}{D+N},$$

which implies the validity of Proposition 3.

The quantity $E_{\Theta|s}\rho_{\mathrm{KL}}(\Theta|s)$ is called a Bayesian estimate for the Kullback–Leibler distance between the above-mentioned distributions (note that an a posteriori mathematical expectation is an optimal Bayesian estimate of a function of a random parameter for a quadratic loss function [13]). We will use the Bayesian estimate obtained (with opposite sign) as a quality criterion for grouping.

We can make the following remarks regarding the practical application of the criterion. For larger dimension of the space of variables, $\ln V^{(0)} \approx 0$; therefore, up to constants, the criterion takes the form

$$Q_{\mathrm{BKL}}(G) = \sum_{k=1}^{K} q^{(k)}(\ln V^{(k)} - \psi(d^{(k)} + n^{(k)} + 1)).$$

The minimization of the criterion leads to a certain compromise between two tendencies: to form taxons of minimum volume and to obtain clusters that contain as many objects as possible.

In this paper, we apply a recursive algorithm to construct a decision tree, in which the Q_{BKL} criterion is used in place of the regularizing criterion (the search scheme remains the same). Moreover, for comparison, we consider a similar algorithm in which a decision is constructed with regard to the criterion Q_{FKL}.

3. ANALYSIS WITH THE USE OF STATISTICAL MODELING

To analyze the algorithm developed, we carried out statistical modeling and repeatedly solved various types of cluster analysis problems. Each type of problems is characterized by properties such as

—the number of classes K,

—the sample volume $n^{(k)}$ for each class,

—the dimension m of the space,

—the number m_q of qualitative variables,

—the set of values of each variable,

—the form of the basic distribution.

The basic distribution for each class is chosen to be multidimensional normal with the same covariance matrix Σ. The vector of mathematical expectations for each class is chosen randomly from the set of integer values of the variables (so that these values for different classes do not coincide). For qualitative variables, the values of the realizations obtained are rounded to the nearest integer. The covariance matrix Σ is defined by two parameters: the value of the diagonal elements σ and the value of off-diagonal elements σ'.

To assess the accuracy of the algorithm, we apply a multiple procedure consisting of the following steps:

—generation of various types of problems with given properties;

—obtaining random samples according to the assigned type of problem;

—construction, by the algorithm analyzed, a group solution for each sample (naturally, the true numbers of classes are not communicated to the algorithm);

—finding an accuracy index averaged over all samples, as well as the corresponding confidence interval.

The accuracy of classification is determined by the Rand index (IR), which represents the relative number of pairs of objects that have either identical or different numbers of classes in the obtained and true classifications (the value of index close to 1 provides evidence for a good consistency of classifications). At the output of the Monte Carlo modeling algorithm, one has accuracy estimates for the algorithm for the types of problems considered.

We analyzed the behavior of the algorithm for constructing a tree for the three quality criteria described above: the regularizing criterion Q_R (the parameter $\alpha = 2$), the criterion based on the frequency estimate for the Kullback–Leibler distance Q_{FKL}, and the criterion involving a Bayesian estimate Q_{BKL} with $d = 1$.

Consider the following example. We generated various types of problems for the case of two classes. The sample size for the first class was 25 and for the second, varied from 10 to 25 with a step of 5; the dimension of the space varied from 5 to 15 with a step of 5; the number of qualitative (Boolean in the case in question) variables was defined randomly; the parameter σ belonged to the set {0.1; 0.2; 0.3; 0.4; 0.5}; and the parameter $\sigma' = 0.5\sigma$. Thus, we considered 60 types of problems altogether. For each type of problems, the modeling procedure was repeated 40 times. It turned out that the number of problems for which some of the algorithms yielded substantially more accurate results compared with other algorithms considered was as follows: 0 for the algorithm based on the criterion Q_R, 0 for Q_{FKL}, and 13 for Q_{BKL}. An example of a graph of the averaged Rand index as a function of the parameter σ is shown in Fig. 2 (the number of simulated samples is 200).

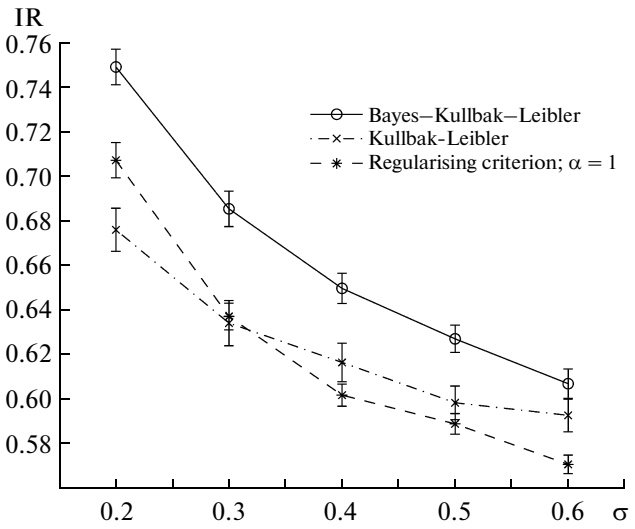

Fig. 2. Example of simulation results ($K = 2$, $n^{(1)} = n^{(2)} = 25$, $m = 10$, $m_q = 5$, and $\sigma' = 0.5$).

Thus, the results of modeling allow us to conclude that, for the types of problems considered, the algorithm based on the Bayesian estimate for the Kullback–Leibler distance much more frequently yielded more accurate results compared with similar algorithms based on the frequency estimate for this distance and on the regularizing criterion.

CONCLUSIONS

We have considered algorithms for constructing taxonomic decision trees that allow one to perform grouping in a space of heterogeneous variables and to form logical classification rules for new objects. We have introduced a modified quality criterion for a tree that is based on the Bayesian estimate for the Kullback–Leibler distance between a distribution corresponding to the clusters formed and a uniform distribution. To this end, we obtained expressions for the expected entropy of a priori and a posteriori distributions of the frequencies of classes. Using statistical modeling, we have shown that there exist examples of problems in which the algorithm based on the Bayesian estimate gives a substantially higher accuracy of classification compared with a similar algorithm using a frequency estimate for the Kullback–Leibler dis-

tance and an algorithm based on a regularizing criterion.

As promising directions of further investigations, we can point out the development of quality criteria for decision trees on the basis of modifications of the Bayesian model that involve various additional assumptions on the classification problem [9]; the development of more efficient search schemes; and the construction of an ensemble of taxonomic decision trees. We are going to compare various algorithms for the cluster analysis of heterogeneous data by means of statistical modeling and by solving applied problems.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 11-07-00346.

REFERENCES

1. Zhexue Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values," in *Proc. 1st Pacific–Asia Conf. on Knowledge Discovery and Data Mining* (Singapore, 1997), pp. 21–34.
2. K. Blekas and A. Likas, "Incremental Mixture Learning for Clustering Discrete Data," in *Artificial Intelligence: Theories, Models and Applications,* Ed. by J. Darzenta, et al. (Springer, Heidelberg, 2004), pp. 210–219.
3. A. S. Biryukov, V. V. Ryazanov, and A. S. Shmakov, "Solving Clusterization Problems Using Groups of Algorithms," Comp. Math. Math. Phys. **48** (1), 168–183 (2008).
4. A. Strehl and J. Ghosh, "Clustering Ensembles—a Knowledge Reuse Framework for Combining Multiple Partitions," J. Mach. Learn. Res. **3**, 583–617 (2002).
5. V. B. Berikov, "The Way to Create the Decision Trees Ensemble in Cluster Analysis," Vychisl. Tekhnol. **15** (1), 40–52 (2010) [in Russian].
6. G. S. Lbov and T. M. Pestunova, "Pooling Objects in the Space of Different Variables," in *Analysis of Nonnumerical Information in Sociological Researches* (Nauka, Moscow, 1985), pp. 141–149 [in Russian].
7. T. Shi and S. Horvath, "Learning with Random Forest Predictors," J. Comput. Graph. Stat. **15** (1), 118–138 (2006).
8. H. Blockeel, L. Raedt, and J. Ramon, "Top–Down Induction of Clustering Trees," in *Proc. 15th Int. Conf. Machine Learning,* Ed. by J. Shavlik (Morgan Kaufmann, 1998), pp. 55–63.
9. G. S. Lbov and V. B. Berikov, *Decision Functions Stability in the Problems of Pattern Recognition and Multitype Information Analysis* (Izd. Inst. Matematiki, Novosibirsk, 2005) [in Russian].
10. N. G. Zagoruiko, *Recognition Methods and Their Application* (Sovetskoe radio, Moscow, 1972) [in Russian].
11. G. M. Fikhtengol'ts, *Course of Differential and Integration Calculus* (Fizmatlit, Moscow, 1960), Vol. 3 [in Russian].
12. I. S. Gradshtein and I. M. Ryzhik, *Tables of Integrals, Sums, Series and Products* (Gos. izd. fiz.–mat. lit., Moscow, 1963) [in Russian].
13. E. L. Lehman, *Theory of Point Estimation* (John Wiley, New York, 1983; Nauka, Moscow, 1991).

Vladimir Borisovich Berikov. Born 1964. Graduated from the Novosibirsk State University in 1986. Received candidate's degree in 1996 and doctoral degree in 2007. Currently is a leading researcher at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. Scientific interests: mathematical theory of pattern recognition and cluster analysis; image analysis; and application of mathematical methods in biology, medicine, and historical studies.