STABILITY AS AN ESTIMATE OF THE DEGREE OF SUBSTANTIATION OF HYPOTHESES DERIVED ON THE BASIS OF OPERATIONAL SIMILARITY

S. O. Kuznetsov

Nauchno-Tekhnicheskaya Informatsiya, Seriya 2
Vol. 24, No. 12, pp. 21-29, 1990

UDC 519.718

The stability of idempotent commutative and associative operations of generalization is investigated. Stability is defined as the good reproducibility on generalization subsamples of the set of all facts from a sample. The concepts introduced in the paper are related to the basic ideas of certain mathematical and applied methods (#P-completeness) including the jackknife method. The difficult solubility of the problem of calculation of stability indexes is established. The boundaries of the variation of these indexes are determined for the sample complementation process. An algorithm computing stability indexes is presented. It is linear relative to the index magnitude. An approximate algorithm, which is polynomial with respect to the length of the input string is also suggested. A computer experiment is described, which used stability indexes for selection of hypotheses in a technical diagnostic problem.

1. SIMILARITY OPERATION AND HYPOTHESES BASED ON IT

Most systems of machine learning (hereafter, ML) operate with a notion of similarity as a means for isolating regular patterns in observable objects. Similarities are also used to classify new objects with the aid of newly found patterns. Similarity is defined either as a relation [1, 2], or metrically [1, 3, 4], or as an operation that puts into correspondence to several initial objects a subobject that expresses their similarity [5-14]. In this paper, similarity is understood as an idempotent commutative and associative operation on pairs of objects, in terms of the theory developed in [7, 8]. These natural properties of a similarity operation allow expressing unambiguously the similarities of sets of objects in terms of pairwise similarities independent of the arrangement of the objects in a database* (e.g., [15]). This definition of similarity is adopted, in particular, in the JSM-method of automatic hypothesis generation (JSM-AHG) [6].

Let S be a set of objects representing a certain object domain. An operation - on pairs of objects from S is called a similarity operation if it specifies on the set S a lower semilattice, i.e., for arbitrary objects x, y, and z from S relations (1)-(4) take place:

(1) $x \sqcap x = x$;

(2) $x \sqcap y = y \sqcap x$;

(3) $x \sqcap (y \sqcap z) = (x \sqcap y) \sqcap z$;

(4) $x \sqcap s_0 = s_0$ for a certain s_0 from S, which is called the empty object.

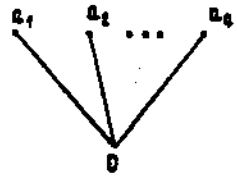
The operation $\sqcap$ defines in a natural fashion on S the relation of embedding $x \sqsubseteq y \Leftrightarrow x \sqcap y = x$ and strict embedding $x \sqsubset y \Leftrightarrow (x \sqsubseteq y) \& (x \neq y)$.

The following examples of operations have properties (1)-(4).

1. The Boolean lower semilattice $\langle 2^{\mathfrak{A}}, \cap, \varnothing \rangle$. This definition of similarity corresponds to a representation of data by a set of descriptors. The operation of similarity is equivalent to the operation of intersection of sets. Representations of this kind are used in many ML systems and, in particular, in JSM-AHG. It is readily obtained by conversion of a representation of similarity by semilattices on n-tuples of a fixed length with

*An alternative to this operation is a parametric family of n-place operations, where n is the number of objects for which similarity is sought (e.g., [11]).

component-by-component "fanlike" specification of similarity (see [14] and also §6 below):



2. A semilattice on N-sets of hypergraphs with ordered labels of nodes and hyperages [12-14], where the result of the similarity operation acting on a pair of sets of hyperedges $\mathcal{G}$ and $\mathcal{H}$ is the set of all embedding-maximal common subhypergraphs of the hypergraphs from $\mathcal{G}$ and $\mathcal{H}$.

3. Interpolational semilattice of intervals [15]. Let (i, j) and (k, l), where $U<i, j, k, l<S$, $i \leqslant j$, $k \leqslant l$, be pairs of numbers indicating boundaries of intervals; U and S are minimal and maximal possible values. Now, the similarity of these intervals is $(i, j) \wedge (k, l) = (p, q)$, where $p = \min(i, k)$, $q = \max(j, l)$. It can readily be verified that the triplet $<\{(i, j) U \leqslant i \leqslant j < S\}, \wedge, (U, S)>$ is a lower semilattice, i.e., it has properties (1)-(4). The initial values of numeric features can be specified by pairs of the form (x, x) if the data are presumed to be precise; otherwise, they can be specified by pairs of different values of this form: lower bound of possible value, upper bound of possible value. We now proceed to definition of hypotheses in accordance with [6]. Suppose that we examine a certain property W of objects from S. The set of all objects from S, of which it is known that they have the property W, will be denoted by S^+; the sets of objects of which it is known that they do not have the property W will be denoted by S^-; the set of objects from S for which it is unknown whether or not they have the property W will be denoted by S^τ. Thus, $S^\tau = S \setminus (S^+ \cup S^-)$.

Definition 1.1. h is local similarity of $X_1, ..., X_n$ from S if $X_1 \sqcap \ldots \sqcap X_n = h$.

Definition 1.2. $<h, \{X_1, \ldots, X_n\}>$ is global similarity with respect to the set $S' \subseteq S$, if $X_1, \ldots, X_n \in S'$, $X_1 \sqcap \ldots \sqcap X_n = h$ and for an arbitrary Y: $Y \in S' \setminus \{X_1, \ldots, X_n\}$ we have $Y \sqcap h \neq h$ (thus h is local similarity of objects from $\{X_1, ..., X_n\}$ and this set is the set of all objects from S' that contain h).

Definition 1.3. $<h, \{X_1, \ldots, X_n\}>$ is positive hypothesis (concerning the cause of the property W) (or (+)-hypothesis) if $<h, \{X_1, \ldots, X_n\}>$ is a global similarity with respect to the set S^+ and h is not a subobject (in the sense of $\sqsubseteq$) of some object from S^-. We will say that h is the head of a hypothesis. Negative hypotheses (or (-)-hypotheses) concerning the cause of the absence of the property W are defined in a dual fashion.

Definition 1.3 is one possible definition of a hypothesis on the basis of the similarity operation. We may require that other conditions be satisfied: stronger, or weaker than, or incomparable in strength with, the condition from definition 1.3 (in JSM-method it is referred to as condition "with prohibition of a counterexample"). This condition is adopted here as one that is typical for machine learning: it requires that a generalization of examples do not include counterexamples as special cases. For the problem of stability of hypotheses, it is essential that hypotheses are global similarities rather than a specific form of the condition.

Hypotheses obtained in conformity with Definition 1.3 can be of an independent practical value, on one hand, and used in the framework of the system for recognition or prediction of the absence/presence of the property W in objects from S^τ on the other. We will formulate an elementary version of a predictive rule: the principle of inductive generalization (PIG) in accordance with [6].

Definition 1.4. An object $P \in S^\tau$ is called (+)-prediction (or (+)-hypothesis of the second kind in the terminology of JSM-method [6]) if there exists (+)- hypothesis $<h, \{X_1, \ldots, X_n\}>$, such that $h \sqsubset P$ and for any (-)-hypothesis $<h', \{Y_1, ... Y_k\}$ we have $h' \not\sqsubset P$.

A negative prediction ((-)-prediction) is defined as dual.

2. STABILITY: MOTIVATION AND PRECEDENTS

Definitions 1.3 and 1.4 presume that the substantive cause of the property W can be the common characteristics possessed by a set of objects $X_1, ..., X_n$ that have the property W. All the characteristics that are dissimilar in these objects are assumed to be immaterial for formation of the hypothesis $<h, \{X_1, \ldots, X_n\}>$. What is the degree of substantiation of this assertion? Obviously, (+)-hypothesis, obtained according to Definition 1.3, can be regarded as better substantiated than a hypothesis that satisfies a weaker condition that requires merely that the head of (+)-hypothesis not be the similarity of (-)-examples ("weak rule") [6].

It is also obvious that a hypothesis corresponding to a global similarity of a larger number of examples is

more substantiated. However, this too is not the extreme situation. In a certain case (e.g., for a hypothesis $H_1 = \langle h_1, \{X_1, \ldots, X_n\} \rangle$), these examples in a certain sense can be "dependent" or "too similar" such as those produced in the same series of experiments by the same experimenter.

In a different situation (e.g., for a hypothesis $H_2 = \langle h_2, \{Y_1, \ldots, Y_n\} \rangle$), examples can be more "independent" such as those obtained by different experimenters or different methods, etc. Thus, they can be dissimilar in all respects, except for the structural fragment h_2. That would mean, in particular, that the hypothesis H_2 can be derived also from a smaller number of examples because the independence of experiments gives hope that the substantial-causative substructures could be separated faster and more reliably from those which are immaterial and merely accompany a certain type of experiments. That, in turn, implies that H_2 will be confirmed on a larger number of subsets of the initial complete set $\{Y_1, \ldots, Y_n\}$ of examples for H_2, i.e., H_2 will have a greater "stability" and "reproducibility" in case of random supplementation or reduction of the set of initial examples. Since each example is obtained in a certain sense at random, the stability to accidents of this kind indicates a higher likelihood of the hypothesis H_2.

The idea of stability has been used in analyses of the reliability of hypotheses of different kinds. Among the methods operating with the notion of stability are the methods of nonparametric statistics - the jackknife method and the bootstrap method [16] - and the methods of sliding control. An elementary example of an implicit application of this notion of stability is construction of extrapolation polynomials. Suppose there are n points $x_1, \ldots, x_n$ of a space R^n. We wish to construct a polynomial from these points such that the points $x_1, \ldots, x_n$ lie on its curve. Generally, it is possible to construct a polynomial of degree not greater than n that satisfies these conditions. However, if it is possible to construct a polynomial P from a certain subset of points $\{x_{i_1}, \ldots, x_{i_k}\} \subset \{x_1, \ldots, x_n\}$ in such a way that all the points $x_1, \ldots, x_n$ lie on its curve, such a polynomial will be of degree not higher than k. Thus, P, as a hypothesis of a regularity will be simpler and, therefore, more reliable than the hypothesis for which the polynomial can only be constructed from the entire set $\{x_1, \ldots, x_n\}$.

As mentioned above, the idea of stability is fundamental to certain nonparametric statistical methods. In particular, in the jackknife method the variances of arbitrary statistics (functions of samples) are estimated in the following fashion. From an initial sample of size n all possible subsamples of size n - 1 are compiled. For the i-th subsample we calculate the value S_i of the statistic S that we wish to examine. Taking the average S^* of these values, we then calculate the mean of the squares of deviations S_i from S^*. The result (within insignificant arithmetic transformation) gives the estimate of the variance of the statistic S according to the jackknife method. The method sometimes yields a better estimate than conventional techniques. Modifications of this method are also possible that make use of all subsamples of size n - 2, n - 3, etc. However, they a require considerable amount of computations.

The bootstrap method evaluates the variance of a statistic in a similar fashion. However, new samples, which are also of size n, are generated from the initial sample by n-fold selection with replacement (each element of the initial sample can appear from zero to n times in the new sample).

Some authors operate with the idea of stability outside the framework of probability models. In particular in [7], the following problem is considered. A set X_N is given, which consists of N points; the values of the function $f(x)$ for these points are known. The objective is to isolate, from a given class of models (functions with parameter θ) $K = \{\mu(x, \theta) : \theta = (\theta_1, \ldots, \theta_m \in R^m\}$ such a model, which on the set $X \subset R^n (X \cap X_N = \varnothing)$ for whose points we must evaluate the function f - has the greatest stability. Stability is defined as proximity for $x \in X$ of the values of the function of the form $\mu(x, \theta)$ constructed on subsamples of the sample X_N to the value of the function constructed from the entire sample X_N. The results of experiments reported in [17] indicate a high selectivity of this method.

Ideas related to the notion of stability are also basic to certain probabilistic logics [18]. Studies in this field, probably initiated by Carnap's work on inductive logic [19], base evaluation of substantiation on the number of universes (or, in a more general statement, on a numeric measure of universes) in which a statement is derivable.

In the above-listed areas, the notion of stability as reproducibility of results on subsamples was realized for numeric data. When no numeric data form the lower sublattices similar to sublattices 1 and 2 from §1, substantiation of the stability of similarity acquires an additional aspect. A good reproducibility of similarity h on the subsets of the set $\{X_1, \ldots, X_n$ (i.e., the existence of a large number of subsets $\{X_{i_1}, \ldots X_{i_k}\}$, such that $X_{i_1} \sqcap \ldots \sqcap X_{i_k} = h$) means that sets from $\mathfrak{M} = \{(X_1 \setminus h), \ldots, (X_n \setminus h)\}$ are "poorly" similar to one another.*

The fact that the similarity of "residues" (the sets from $\mathfrak{M}$) is nontrivial ($\neq 0$) supports the idea that the

---

*\ is either Boolean difference for lattice 1 or pseudodifference for lattice 2 [12-14].

cause of W in most likelihood is not h but certain $h_1, \ldots, h_q \supset h$. By virtue of properties (1)-(4) of similarity operation $\sqcap$, similarities of a smaller number of objects are objects that are not smaller (in the sense of relation $\sqsubset$).

Thus, the poor reproducibility of hypotheses on subsets of examples or a low stability indicate that the "residues" from the set X contain substructures or parts that are essential for manifestation of the property W, and h can be the cause of the property W only with combined when some "additions" (that complement h to $h_1, \ldots h_q$).

3. BASIC DEFINITIONS

All the definitions will be given for positive hypotheses; for negative hypotheses the definitions of stability are dual. Thus, let $H = \langle h, \{X_1, \ldots, X_n\} \rangle$ be (+)-hypotheses of the first kind. We denote

$$\langle H \rangle_j = \{\{X_{i_1}, \ldots, X_{i_j}\} \mid \{X_{i_1}, \ldots, X_{i_j}\} \subset \{X_1, \ldots, X_n\},$$
$$X_{i_1} \sqcap \ldots \sqcap X_{i_j} = h\};$$
$$\langle H \rangle_\Sigma = \bigcup_{j=2}^{n-1} \langle H \rangle_j;$$
$$g_j(j, H) = \# \langle H \rangle_j, \quad g_\Sigma(\Sigma, H) = \# \langle H \rangle_\Sigma.$$

Wherever it is obvious which hypothesis is discussed, the arguments at $g_j(q, H)$ will be omitted, i.e., we will write g_j (or g_Σ).

Definition 3.1. The stability index of a hypothesis $\langle H, \{X_1, \ldots, X_n\} \rangle$ of j-th level for $2 \leqslant j \leqslant n-1$ is

$$I_j = \frac{g_j}{\binom{n}{j}}.$$

Definition 3.2. The integral stability index of a hypothesis $\langle H, \{X_1, \ldots, X_n\} \rangle$ is

$$I_\Sigma = \frac{g_\Sigma}{2^n - n - 2}.$$

Definition 3.3. The averaged stability index of a hypothesis $\langle H, \{X_1, \ldots, X_n\} \rangle$ is

$$I_m = \frac{1}{n-2} \left(\sum_{j=2}^{n-1} I_j \right).$$

Stability indices are connected with the similarity operation in the same fashion as the sample average is connected with a sample variance (calculated by the jackknife method; see above the expression for the jackknife variance estimate) in statistics. It should be clear that a bootstrap definition of stability estimates is unjustified. Indeed, taking i times, in a bootstrap sample of size n, a certain object from $\{X_1, \ldots, X_n\}$ would be equivalent, by virtue of idempotency (property 1) of the operation $\sqcap$, of merely reducing the number of other objects in a bootstrap sample. As a result, we would obtain the same types of samples as with elimination of examples but these samples would make different contributions to stability indices. Samples with a large (~n) and small (~2) number of examples would have less probability of occurrence in new samples than samples with the average number of examples. This preference is unjustified because there are no samples a priori preferable.

The following property of stability indices is a corollary of the elementary property of monotone Boolean functions*: the relative number of units of a monotone Boolean function (J + 1)th layer of a Boolean hypercube

---

*For a fixed hypothesis $H = \langle h, \{X_1, \ldots, X_n\} \rangle$, it is the function.

$$\sum_{Y \sqsubset \{X_1, \ldots, X_n\}}^{f(Y)} =$$
$$= \begin{cases} 1, & \text{if } Y = \{X_{i_1}, \ldots, X_{i_j}\} \text{ and } X_{i_1} \sqcap \ldots \sqcap X_{i_j} = h; \\ 0, & \text{if } Y = \{X_{i_1}, \ldots, X_{i_j}\} \text{ and } X_{i_1} \sqcap \ldots \sqcap X_{i_j} \neq h. \end{cases}$$

is greater than in the jth layer.

Lemma 3.1. For an arbitrary hypothesis $\langle n, \{X_1, \ldots, X_n\}\rangle$ we have

$$I_2 \leqslant \ldots \leqslant I_{n-1}.$$

Proof. We consider families $\langle H\rangle_j$, $\langle H\rangle_{j+1}$, and a bipartite graph B constituted by the layers j and j + 1 of the Boolean hypercube $2^{\{X_1,\ldots,X_n\}}$. In the graph B, each of $\binom{n}{j+1}$ nodes of the layer (j + 1) is connected with the node (j + 1) of the jth layer; each of $\binom{n}{j}$ nodes of the j-th layer is connected with n - j nodes of the (j + 1)th layer. We isolate in the graph B those nodes that correspond to the families $<H>_j$ and $<H>_{j+1}$. Since any superset of size j + 1 of sets from $<H>_j$ is a set from $<H>_{j+1}$, the number of edges in the graph that connect nodes corresponding to sets from $<H>_j$ with the nodes corresponding to sets from $<H>_{j+1}$ is $e = g_j (n - j)$. On the other hand, generally, not each subset of size j of a set from $<H>_{j+1}$ is a set from $<H>_j$. Therefore, the number of edges e is not greater than $g_{j+1}(j+1)$. Hence, $g_j(n-j) \leqslant g_{j+1}(j+1)$. Since $\dfrac{\binom{n}{j+1}}{\binom{n}{j}} = \dfrac{n-j}{j+1}$, therefore, $\dfrac{g_j}{g_{j+1}} < \dfrac{j+1}{n-j} = \dfrac{\binom{n}{j}}{\binom{n}{j+1}}$ and $I_j = \dfrac{g_j}{\binom{n}{j}} < \dfrac{g_{j+1}}{\binom{n}{j+1}} = I_{j+1}$. □

4. STABILITY VARIATION CAUSED BY AN EXPANSION OF A SET OF EXAMPLES

Let the set of initial examples S be expanded by inclusion of a new example E.

Definition 4.1. A new (-)-example E refutes (+) hypothesis $<h, \{X_1, \ldots X_n\}>$, if from the set of examples S^+, $S^- \cup \{E\}$ one cannot derive (+)-hypothesis with head h (i.e., $h \sqsubseteq E$).

Definition 4.2. A new (+)-example E confirms (+)-hypothesis $<h, \{X_1, \ldots, X_n\}>$ if $h \sqsubseteq E$.

Stability indices of a hypothesis that has the head h after introduction of new k examples will be denoted by superscript k, e.g., I_j^k. For convenience of notation, we will also set

$$I_n = 1,$$
$$I_j = 0 \quad \text{for} \quad j \in Z \setminus \{2, \ldots, n\}.$$

Theorem 4.1. Suppose that a set of examples has been expanded to include k new (+)-examples confirming a hypothesis $<h, \{X_1, \ldots, X_n\}>$. The stability indices of a hypothesis with head h then lie in the following limits

$$\frac{1}{\binom{n+k}{j}}\left[g_j + \binom{k}{1} g_{j-1} + \cdots \right.$$
$$\left. \cdots + \binom{k}{k-1} g_{j-k+1} + g_{j-k}\right] < I_j^k <$$
$$< \frac{1}{\binom{n+k}{j}}\left[g_j + \binom{n}{j-1} + \cdots + \binom{n+k-1}{j-1}\right]$$

at $(2 \leq j \leq n + k - 1)$;

$$\frac{2^k \cdot g_\Sigma + 2^k - 1}{2^{n+k} - (n+k+2)} < I_\Sigma^k < \frac{g_\Sigma + 2^n (n^k - 1) - k}{2^{n+k} - (n+k+2)}.$$

Proof. 1. The lower bounds for indices I_j^k (or I_j^k). $j \in \{2, \ldots, n+k-1\}$. At a given initial value of the index I_j, the value I_j^k is minimal if any subset of the set $\bar{K}$ of new confirming examples yields h in the intersection of only those sets of examples from $\{X_1, \ldots, X_n\}$, which themselves have h in the intersection. In more rigorous

terms $I_j^k = \frac{g_j^k}{\binom{n+k}{j}}$ takes the minimal value if for any p, q $1<p<k$. $1<q<n$. $\tilde{X}_{i_1}\sqcap\ldots\sqcap\tilde{X}_{i_p}\sqcap(\tilde{X}_{i_1}\sqcap\ldots\sqcap\tilde{X}_{i_q})=h$ if and only if $X_{j_1}\sqcap\ldots\sqcap X_{j_q}=h$. Here, $\tilde{X}_{i_1},\ldots,\tilde{X}_{i_p}\in\{\tilde{X}_1,\ldots,\tilde{X}_k\}$. $\{\tilde{X}_1,\ldots,\tilde{X}_k\}$ is the set of new examples $X_{i_1},\ldots,X_{i_q}\in\{X_1,\ldots,X_n\}$.

Let us consider the terms that make up the value of g_j^k. The first term, i.e., g_j corresponds to subsets of the size of the j-th set of initial examples $\{X_1, \ldots, X_n\}$. The other terms appear as $g_{j-s}\binom{k}{s}$. They are obtained by virtue of the fact that an arbitrary set $\{X_{i_1},\ldots,X_{i_{j-s}}\}$ from $<H>_{j-s}$ can be supplemented with any new confirming examples $\tilde{X}_{r_1},\ldots,\tilde{X}_{r_k}$ to obtain a set $\{X_{i_1},\ldots,X_{i_{j-s}},X_{r_1},\ldots,X_{r_k}\}$, which also has h in the intersection of all its elements $X_{i_1}\sqcap\ldots\sqcap X_{i_{j-s}}\sqcap\tilde{X}_{r_1}\sqcap\ldots\sqcap X_{r_k}=h$. Therefore, $g_j^k=\sum_{s=0}^{k} g_{j-s}\cdot\binom{k}{s}$ and the lower estimate for I_j^k has been proved.

2. The lower bounds for the index I_Σ (or $\underline{I}_\Sigma$). When a new confirming example is received, each value of g_i^1, $2\leqslant i\leqslant n-1$ becomes not less than the value indicated in the preceding section. The entire sequence of values of g_i^1, is $g_2, g_2+g_2, \ldots, g_{n-2}+g_{n-1}, g_{n-1}+1$. Their sum is $g_\Sigma=g_2+(g_2+g_2)+\ldots+(g_{n-2}+g_{n-1})+(g_{n-1}+1)=2g_\Sigma+1$. Therefore, after receiving k new confirming examples, the value of g_Σ is not less than $2^k\cdot g_\Sigma+2^k-1$ (i.e., $g_\Sigma^k\leqslant 2^k\cdot g_\Sigma+2^k-1$) and

$$\underline{I}_\Sigma^k>\frac{2^k\cdot g_\Sigma+2^k-1}{2^{n+k}-(n+k+2)}.$$

3. The upper bounds for the indices I_j^k (or $\bar{I}_j^k$), the values of $j\in\{2,\ldots,n+k-1\}$ are obtained from analysis of the sequence of new examples $\tilde{X}^1,\ldots,\tilde{X}^k$ that confirm the hypothesis H and are of the following form: $\tilde{X}^i$ in the intersection with any previous example confirming H yields h; more precisely, $\tilde{X}^i\sqcap X=h$ for $X\in\{X_1,\ldots,X_n,\tilde{X}^1,\ldots,\tilde{X}^{i-1}\}$. In that case, for k = 1, g_j^1 is the sum of the number of old examples g_j and the new examples formed from j - 1 old examples and a single new example: $g_j^1=g_j+\binom{n}{j-1}$. Suppose that for $k=t$ $g_j^t=g_j+\binom{n}{j-1}+\ldots+\binom{n+t-1}{j-1}$. In that case, for $k=t+1$ g_j^{t+1} is the sum of the number of old examples g_j^t and new examples formed from j - 1 old examples and a single new one:

$$g_j^{t+1}=g_j^t+\binom{n+t}{j-1}=g_j+\binom{n}{j-1}+\ldots$$
$$\ldots+\binom{n+t-1}{j-1}+\binom{n+t}{j-1}.$$

4. The upper bounds for the index I_Σ (or $\bar{I}_\Sigma^k$). The initial values of g_i were $g_2,\ldots,g_{s-1}, g_s, g_{s+1},\ldots, g_{n-2}, g_{n-1}$; after the arrival of X^1, these values are expressed by $g_2+\binom{n}{1},\ldots,g_s+\binom{n}{s-1},\ldots,g_{n-1}+\binom{n}{n-2}, g_n+\binom{n}{n-1}$, respectively, where $g_n = 1$. In this case,

$$g_\Sigma^1=(g_2+\ldots+g_{n-1})+\left(\binom{n}{1}+\ldots\right.$$
$$\left.\ldots+\binom{n}{s-1}+\ldots+\binom{n}{n-1}+1\right)=g_\Sigma+2^n-1.$$

Hence $g_\Sigma^k=g_\Sigma+2^{n+k-1}+\ldots+2^n-k=g_\Sigma+2^n(2^k-1)-k$ and $\bar{I}_\Sigma^k=\frac{g_\Sigma+2^n(2^k-1)-k}{2^{n+k}-(n+k+2)}$. □

Are the estimates indicated in the theorem exact? We will indicate sequences for which these estimates are exact on a fairly large number of new examples k.

We consider the Boolean case: $S=2^U$, $U=\{d_1,\ldots,d_n\}$. Let $\langle h, \{X_1,\ldots,X_n\}\rangle$ be an arbitrary (+)-hypothesis. Suppose that $X^1=X_1\setminus\{h\}$, $U^1=X_1^1\cup\ldots\cup X_n^1$, $U^2=U\setminus(U^1\cup n)$, $U^2=\{d_{i_1},\ldots,d_{i_k}\}$.

A sequence of new examples realizing the upper bound of the variation of I_Σ is constructed as follows: the pth new example is of the form $X_p=h\cup\{di_p\}$. Obviously, any pair of new examples and any new example with any old example forms h at the intersection. The maximum number of such new examples is $k=|U^2|=|U|-|U^1|-|h|$.

A sequence of new examples realizing the lower bounds is constructed similarly: the pth new example is of the form $X_p=h\cup U^1\cup\{di_p\}$. As can readily be seen, new examples give h at the intersection with only those old examples, which themselves have h at the intersection, and, in this fashion, the lower bounds of stability indices

indicated in Theorem 4.1 are realized. The maximum number of new examples k which realize precisely the lower bound is also $k=|U^2|=|U|-|U^1|-|h|$.

We will now examine the limits of the lower and upper bounds of stability indices for $k \to \infty$.

The lower limits of the level indices behave differently: for the indices of the upper levels, they approach 1 in the limit; for the lower levels, they approach 0. Indeed, by virtue of Theorem 4.1, we have $\underline{I}_{n+k-1}=\frac{1}{\binom{n+k}{n+k-1}}\left(g_{n-1}+\binom{k}{k-1}\right)$ and $\lim\limits_{k\to\infty}\underline{I}_{n+k-1}=1$. On the other hand, $\underline{I}_2=\frac{1}{\binom{n+k}{2}}\cdot g_2$ and $\lim\limits_{k\to\infty}\underline{I}_2=0$.

The behavior of the lower limits of the indices of the middle levels and the averaged index I_m^k remains unclear.

The limit of the lower bound of the integral stability index is strictly greater than zero and smaller than one. Indeed, the function $f(x)=2^n-\frac{n+x+2}{2^x}$ increases monotonically at $x > 0$ because

$$f'(x)=-\frac{-2^x+(n+x+2)\cdot 2^x\cdot\ln 2}{(2^x)^2}>0$$

at $x > 0$ and, therefore, the function $\underline{I}_\Sigma^k(k)$ decreases monotonically, as k grows and approaches $\lim\limits_{k\to\infty}\underline{I}_\Sigma^k(k)=\frac{g_\Sigma+1}{2^n}>0$.

The upper bounds of stability indices behave uniformly: they increase monotonically and, in the limit, approach 1. Indeed,

$$\overline{I}_1^k=\frac{1}{\binom{n+k}{2}}\left[g_1+\binom{n}{1}+\ldots+\binom{n+k-1}{1}\right]=$$

$$=\frac{g_1}{\binom{n+k}{2}}+\frac{\binom{n+k}{2}-\binom{n}{2}}{\binom{n+k}{2}}\to 1$$

and, by virtue of Theorem 3.1, it is also true for $\overline{I}_1^k, \ldots, \overline{I}_{n+k-1}^k, \overline{I}_m^k$.

$$\overline{I}_\Sigma^k(k)=\frac{2^{n+k}-k-(2^n-g_\Sigma)}{2^{n+k}-(n+k+2)}=$$

$$=\frac{f(k)-d}{f(k)-b}=\left(1-\frac{d-b}{f(k)-b}\right),$$

where $f(k)=2^{n+k}-k$ is a strictly monotonically increasing function for $k \geq 0$ and $d=(2^n-g_\Sigma)\geq(n+2)=b$ because $g_\Sigma<2^n-n-2$. $f(k)-b=2^{n+k}-k-n+2>0$ for any k, n > 0, d−b = const. Therefore, $\overline{I}_\Sigma(k)$ increases strictly monotonically and $\lim\limits_{k\to\infty}\overline{I}_\Sigma(k)=1$.

We should note that the monotone variation of the lower and upper limits of $I_\Sigma^k(k)$ allows us to strengthen the statement of Theorem 4.1 as far as this index is concerned. These limits hold not only for confirming examples but also for any nonrefuting examples. Indeed, a new (-)-example cannot change the stability of a hypothesis - it can only refute the hypothesis itself. (+)-Examples that do not confirm a (+)-hypothesis <h, {X_1, ..., X_n}> by virtue of this monotonicity cannot cause an integral stability index to be smaller than the lower bound or larger than the upper bound.

An analysis of the asymptotic behavior makes it possible to formulate hypotheses concerning the behavior of the index I_Σ alone: in most likelihood, with the reception of new examples, the value of this index will increase because there is practically no room for it to decrease any further (Fig. 1).

For example, for the hypothesis <h, {X_1, ..., X_n}>, $n=4$, the ratio of the initial value of I_Σ^0 to the limit $\lim\limits_{k\to\infty}\underline{I}_\Sigma^k$ for the average (~1/2) value of I_Σ^0 is

$$\frac{g_\Sigma}{2^4-4-2}\cdot\frac{2^4}{g_\Sigma+1}\approx 1.3.$$

At n = 8,

$$\frac{g_\Sigma}{2^8-8-2}\cdot\frac{2^8}{g_\Sigma+1}\approx 1.04.$$

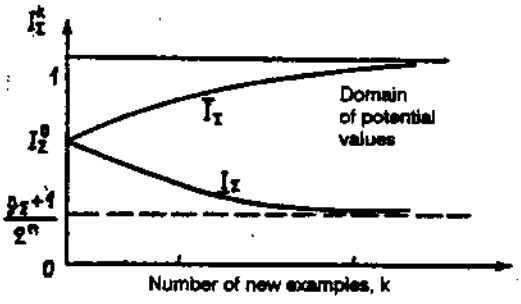


Fig. 1

Thus, the difference is very small as compared with the potential doubling of $\lim_{k\to\infty} I_\Sigma^k$ relative to I_Σ^0. Now, if we represent the set of examples of a hypothesis $<h, \{X_1, ..., X_n\}>$ as produced by supplementation of a certain initial set of size $r < n$, we can conclude that for hypotheses with larger n I_Σ is more likely to be greater than for hypotheses with small m. This "soft" correlation between I_Σ and a number of examples allows us to give preference to the integral stability index as the most representative of the indices for which the asymptotic behavior of the lower bounds is known: on one hand, I_Σ explicitly reflects the stability of a hypothesis; on the other hand, implicitly it reflects a number of confirming examples.

5. ALGORITHMIC COMPLEXITY OF COMPUTATION OF STABILITY INDICES

Regrettably, exact calculation of stability indices in all likelihood (if P ≠ #P), even when the examples are represented by sets, cannot be accomplished within time that is polynomial with respect to the size of the hypothesis $<h, \{X_1, ..., X_n\}>$, i.e., $|h| + n$. This can be established by virtue of Theorem 5.1. We recall that #P [20] is the class of problems that can be calculated by a nondeterministic Turing's machine within polynomial time. A problem for #P is said to be #P-complete if any problem from #P is reducible to it (after Turing). The set of #P-complete problems comprises not only enumerative problems that correspond to NP-complete recognition problems (the problems of the number of Hamiltonian paths in a graph, the number of ensembles implementing CNF, etc.) but also enumerative problems for which the corresponding recognition problems are polynomially soluble (e.g., the problem of the number of maximum paired combinations).

Theorem 5.1. Suppose that a Boolean algebra is used for data representation (see §1). In that case, the problem SI_Σ of calculation of the stability index I_Σ of an arbitrary hypothesis $<h, \{X_1, ..., X_n\}>$ is #P-complete.

For the proof we will introduce certain auxiliary problems.

The problem of the number of node coverings (NNC)

Given: Graph $G = <V, E>$.

Find: The number of node coverings, i.e., $\#\{V' \subseteq V \mid$ if $(u, v) \in E$, then $u \in V'$ or $v \in V'\}$.

The problems of the number of implicants (NI)

Given: Monotonous 2-CNF, i.e., the formula $F = C_1 \wedge \ldots \wedge C_r$, where

$$C_l = (x_{l_1} \vee x_{l_2}) x_{l_1}, x_{l_2} \in X = \{x_1, \ldots, x_n\}.$$

Find: $\#P\{Y \mid Y \subseteq X, \bigwedge_{x_j \in Y} x_j \to F\}$.

Problem of the number of subfamilies with fixed intersection (NSFI).

Given: A finite set $\mathcal{U}$ and $\mathfrak{B} \subset 2^{\mathcal{U}}$, a family of different sets $\mathfrak{B} = \{X_1, \ldots, X_k\}$, where $X_1 \cap \ldots \cap X_k = h$.

Find: The number of subfamilies $\mathfrak{B}'$ of the family $\mathfrak{B}$ which are such that the intersection of all members of the subfamily $\mathfrak{B}'$ is h, i.e.,

$$\#\{\mathfrak{B}' \subseteq \mathfrak{B} \mid \mathfrak{B}' = \{X_{t_1}, \ldots, X_{t_s}\} \text{ and } X_{t_1} \cap \ldots \cap X_{t_s} = h\}.$$

We will now prove lemmas that lead to Theorem 5.1.

Lemma 5.2. The NI problem is #P-complete.

Proof. We will reduce to the NI problem the following problem "the number of Boolean sets that satisfy 2-CNF $F = C_1 \wedge C_2 \wedge \ldots \wedge C_r$», where $C_l = (y_{l_1} \vee y_{l_2})$ and $y_{l_j} \in X$». The #P-completeness of this problem has

been proved in [20]. Suppose that $|X| = n$ and $A = (a_1, ..., a_n)$ is a Boolean ensemble that satisfies F. Let $\{j_1, ..., j_k\}$ be serial numbers of single components of the Boolean ensemble A. We form a conjunction $Y_j = y_{j_1} \wedge \ldots \wedge y_{j_k}$ where $y_{j_i} \in X$. Obviously, Y_j is the implicant of two-CNF F. Conversely, each implicant $Y_m = y_{m_1} \wedge \ldots \wedge y_{m_s}$ of the formula F has a corresponding Boolean ensemble $A^m = (a_1^m, \ldots, a_n^m)$, where positions $m_1, ..., m_s$ are one-positions and the remaining positions are zero-positions; the Boolean ensemble A^m fulfills F. The reducibility has been proved. Since the notation of NI cannot exceed the size $\log 2^{|X|} = X$ and reducibility is polynomial, Lemma 5.2 has been proved.

Lemma 5.3. The NNC problem is #P-complete.

Proof. The membership of the NNC problem in the class #P is not questioned, because the size of notation of the problem solution is not greater than $\log 2^{|V|} = |V|$ bits. We will prove the #P-completeness of the problem by reducing to it #P-complete problem NI. As in [21], we construct from an arbitrary 2-CNF F a graph $G = \langle V, E \rangle$ and $V = \{u_1, ..., u_k\}$ (where each u_i corresponds to a variable x_i from F) and $E = \{(u_i, u_j) \mid (x_i \vee x_j)$ is included in the conjunction$\}$. Any nodal covering of G corresponds to the implicant of F, and conversely; any implicant of F corresponds to a certain nodal covering of G. The reducibility is implemented within a time linear with respect to the size of F □.

Lemma 5.4. The NSFI problem is #P-complete.

Proof. In Turing's terms we will reduce the NNC problem to a special case of this problem (at $h = \varnothing$). Suppose that we have an arbitrary graph $G = \langle V, E \rangle$ with no isolated nodes (which does not impair the generality of the analysis), where $V = \{a_1, \ldots, a_n\}$, $E \subseteq V \times V$. We denote by N(V) the set of edges from E which are incident to the node v; $\bar{N}_E(V) = E \setminus N(v)$. For certain v_i, $v_j \in V$, we have $N(v_i) = N(v_j) = \varnothing$ if and only if $N(v_i) = N(v_j) = \{(v_i, v_j)\}$, i.e., v_i and v_j are incident to one and the same edge, which is not connected with other nodes of the graph G. We will calculate the NNC of the graph as follows: we isolate in the graph G isolated edges (i.e., edges of the form $(v_i, v_j) \in E$, for which $N(v_i) = N(v_j)$). Suppose that there are k such edges in the graph G. The nodal covering of these edges can be executed in 3^k ways (three ways per each edge: it can be covered by either of the two nodes or by both). If the remaining edges of the graph G can be covered in d various ways, then all the edges of the graph have $d \cdot 3^k$ coverings. We have only to determine the number of nodal coverings of unisolated edges of the graph (we will denote this set by E'). Suppose that edges from E have incident nodes from the set $V' \subseteq V$. Now, on any two different nodes v_i, $v_j \in V'$, we have $N(v_i) \neq N(v_j)$. By definition, the nodes $v_1, \ldots, v_r \in V'$ form a covering of E if and only if $N(v_1) \cup \ldots \cup N(v_r) = E'$ or, by de Morgan's law, $\bar{N}_{E'}(v_1) \cap \ldots \cap \bar{N}_{E'}(v_r) = \varnothing$. Thus, the set $\{v_1, ..., v_r\}$ forms the nodal covering of the graph $G = \langle V', E' \rangle$ if and only if the intersection of all sets of the family $\bar{N}_{E'}(v_1), \ldots, \bar{N}_{E'}(v_r)$ is empty. We have thus reduced the NNC problem for $\langle G = \langle V, E \rangle$ to NSFI problem with $\mathfrak{A} = \{\bar{N}_{E'}(a_1), \ldots, \bar{N}_{E'}(a_n)\}$. The reducibility is polynomial because the dimension of the set $\mathfrak{A}$ is not greater than $n\left(\left(\frac{n}{2}\right) - (n-1)\right)$, i.e., $0(n^3)$.□

Theorem 5.1 is a simple corollary to Lemma 5.4. In conditions where the hypothesis H is not contradictory and the set of all examples generating it is $\mathfrak{A}$, NSFI is g_Σ.

Corollary. The problem of determination of I_j is #P-complete.

Indeed, if we know the values of I_j for $2 \leq j \leq n - 1$, we can calculate the value of I_Σ as

$$I_\Sigma = \frac{\sum_{j=1}^{n-1} I_j \cdot \binom{n}{j}}{2^n - n - 2}.$$

We will describe and evaluate the complexity of the algorithm that computes the stability indices of a hypothesis $\langle h, \{X_1, ..., X_n\} \rangle$ in the Boolean case.

We assume that all examples and hypotheses can be represented by Boolean vectors of size $|U| = m$ and that two bit rows of size b can be multiplied or compared in b computer operations.

Level 1.

1. Generate subfamilies of the family $\{X_1, ..., X_n\}$ of size n - 1, i.e., the family $\mathfrak{F}_1$: $\{X_1, \ldots, X_{n-1}\}$, $\{X_1, \ldots, X_{n-2}, X_n\}, \ldots, \{X_2, \ldots, X_n\}$. Each of the subfamilies can be represented by a set of numbers from 1 to n. This step takes $0(\log n \cdot (n - 1)$ computer operations and $0(\log n \cdot (n - 1) \cdot n)$ memory locations.

2. For each set from the family $\mathfrak{F}_1$, compute the intersection of all its members and match against h. Count the number of intersections that coincide with h. The step takes up $0(n(n-1)m)$computations and $0(n(n-1)m)$

memory locations.

Level i.

The input to the level i is the family $\mathcal{F}_{n-i+1}: \{Y_1, \ldots, Y_{t_{i-1}}\}$ of subfamilies of size n - i + 1 of the family $\{X_1, \ldots, X_n\}$. The members of these subfamilies in the intersection form h (and, therefore, $t_{i-1} = I_{n-i+1}\binom{n}{i-1}$, i.e., $t_j = g_{n-j}$).

1. For each Y_s from $\mathcal{F}_{n-i+1}$ generate n - 1 + 1 subfamilies $Y_s^1, \ldots, Y_s^{n-i+1}$ of size n - i. The families that are generated are represented by lexicographically ordered ensembles of numbers: each X_i from $\{X_1, \ldots, X_n$ is assigned a number i. A subfamily Y_s^r is assigned an ensemble of numbers corresponding to all the sets from Y_s^r.

The generation of the sets Y_s takes up $0(t_{i-1}\cdot\log n\cdot(n-i+1)\cdot n)$ computations. There are at most $t_{i-1}(n-i+)$, such sets. Therefore, the arrangement in the lexicographic order will take up at most $0(t_{i-1}\cdot\log n\cdot(n-i+1)\cdot\log(t_{i-1}(n-i+1)))$ computations. The memory volume necessary does not exceed $0(t_{i-1}(n-i+1)\cdot\log n)$.

2. We take the intersections of all sets in each Y_s^r and compare the results with h. It takes up at most $0(t_{i-1}\cdot(n-i+1)\cdot m\cdot(n-i))$computations, where $t_{i-1}(n-i+1)$, is the maximum number of families Y_s and (n - i) is the maximum size of Y_s.

3. Parallel to intersection and comparison, compute the number of intersections that match h. This will take at most $0(t_{i-1}\cdot(n-i+1)\cdot m)$ computations and $0(\log(t_{i-1}\cdot(n-i+1)m))$ memory locations.

The composite complexity of all steps in the computation of I_j taking into account that $\log t_{i-1} \le \log 2^n = n$, does not exceed $0\left(\left(\sum_{i=1}^{n-j} t_i\right)\cdot m\cdot n^3\right)$ computations and requires at most $0\left(\left(\sum_{i=1}^{n-j} t_i\right)\cdot m\cdot n\cdot\log^2 n\right)$ memory locations. The evaluation of I_Σ, I_m takes up at most $0\left(\left(\sum_{i=1}^{n-1} t_i\right)\cdot m\cdot n^3\right)$ computations and $0\left(\left(\sum_{i=1}^{n-1} t_i\right)\cdot m\cdot n\cdot\log^2 n\right)$ memory locations. Thus, the following statement holds.

Theorem 5.5.

1. The stability index $I_j = \frac{g_j}{\binom{n}{j}}$ $2 < j \le n-1$ can be computed within a time that is linear with respect to $\max_{2<i<j} g_i$.

2. The integral stability index $I_\Sigma = \frac{\sum_{j=2}^{n-2} g_j}{2^n - n - 2} = \frac{g_\Sigma}{2^n - n - 2}$ can be computed within a time that is linear with respect to g_Σ.

This assertion and the determination of #P-completeness of the problems of calculation of stability indices indicates that this algorithm is optimal within a polynomial multiplier $0(n^3)$ (in the sense of the definition of optimality of algorithms for #P-complete problems from [20]).

Effective algorithms can be proposed for computing the lower and upper estimates of integral I_Σ and average I_m stability indices for the Boolean case $(S = \langle 2^{2^l}, \cap, \varnothing\rangle)$ on the basis of the following propositions.

Theorem 5.5. For an arbitrary (+)-hypothesis H = <h, $\{X_1, \ldots, X_n\}$> we have

$$\frac{g_2 + \ldots + g_k}{\binom{n}{2} + \ldots + \binom{n}{k}} < I_\Sigma < \frac{g_{n-r} + \ldots + g_{n-1}}{\binom{n}{n-r} + \ldots + \binom{n}{n-1}}.$$

Theorem 5.6. For an arbitrary (+)-hypothesis H = <h, $\{X_1, \ldots, X_n\}$> we have

$$\frac{1}{k}\left(\frac{g_2}{\binom{n}{2}} + \ldots + \frac{g_k}{\binom{n}{k}}\right) < I_m <$$
$$< \frac{1}{r}\left(\frac{g_{n-r}}{\binom{n}{n-r}} + \ldots + \frac{g_{n-1}}{\binom{n}{n-1}}\right).$$

We will first prove the following lemma.

Lemma 5.7. For arbitrary sequences (a_i), (b_i), which are such that $a_i \ge 0$, $b_i > 0$ and $\frac{a_i}{b_i} < \frac{a_{i+1}}{b_{i+1}}$, and for an arbitrary $s \ge 1$, we have

$$\frac{a_1+\ldots+a_s}{b_1+\ldots+b_s}<\frac{a_1+\ldots+a_{s+1}}{b_1+\ldots+b_{s+1}}<$$
$$<\frac{a_2+\ldots+a_{s+1}}{b_2+\ldots+b_{s+1}}<\frac{a_{s+1}}{b_{s+1}}.$$

The proof of the first inequality will be done by induction with respect to s, s = 1. Suppose that $\frac{a_1}{b_1}<\frac{a_2}{b_2}$. In that case, $\frac{a_1}{b_1}<\frac{a_1+a_2}{b_1+b_2}<\frac{a_2}{b_2}$. Indeed

$$\frac{a_1+a_2}{b_1+b_1}-\frac{a_1}{b_1}=\frac{a_1b_1+a_2b_1-a_1b_1-a_1b_2}{b_1(b_1+b_2)}>0,$$
$$\frac{a_1+a_2}{b_1+b_2}-\frac{a_2}{b_2}=\frac{a_1b_2+a_2b_2-a_2b_1-a_1b_2}{b_2(b_1+b_2)}<0.$$

Suppose that for s > m the statement has been proved and $a_1 + \ldots + a_m = A$, $b_1 + \ldots + b_m = B$. In that case,

$$\frac{a_1+\ldots+a_{m+1}}{b_1+\ldots+b_{m+1}}-\frac{a_1+\ldots+a_m}{b_1+\ldots+b_m}=\frac{A+a_{m+1}}{B+b_{m+1}}-\frac{A}{B}=$$
$$=\frac{AB+a_{m+1}B-BA-b_{m+1}A}{B(B+b_{m+1})}=\frac{a_{m+1}B-b_{m+1}A}{B(B+b_{m+1})}=\Delta.$$

But $\frac{a_{m+1}}{b_{m+1}}>\frac{a_m}{b_m}$ and by virtue of the assumption in the induction $\frac{a_m}{b_m}>\frac{A}{B}$ and the numerator Δ is nonnegative.□

Other inequalities are proved similarly. The proof of Theorem 5.5 follows directly from Lemmas 3.1 and (5.7). Theorem 5.6 is proved similarly. Theorems 5.5 and 5.6 allow us to calculate the upper and lower bounds of I_Σ and I_m within the time $0(|U|\cdot n^{k+r})$.

We know from experience that even for a small number of examples confirming a hypothesis, it may be difficult to calculate the exact value of the indices. On the other hand, computations under Theorems 5.5 and 5.6 do not always yield good approximations: at those values of the parameters k and r from these theorems which make it possible to employ computer resources (see §6), the upper and lower estimates are often far apart. Presumably, an effective computation of good approximations of stability indices would be possible with the aid of Monte Carlo techniques.

6. COMPUTER EXPERIMENT

An experimental analysis of stability indices was conducted in a search by means of the JSM-system of the potential causes of defects in polyamides produced by the Plastmassy Plastics Factory.

The technological and material parameters of polymer production process was characterized by eight elements: characteristics of the weight share of volatile substances, acidity, and polarization of the raw material; the serial number of the autoclave where the polymer was produced; the time of day and the work-shift number during which the polymer was produced; and the time elapsed after the end of degassing until unloading and the idle time.

The defects in the end product were due to deviations from the standard in the following four characteristics of the polymer: conformity with specification, conformity with the All-Union State Standard GOST, absence of non-cut-through areas (a characteristic of viscosity), and absence of foreign inclusions.

The similarity of the descriptions of the industrial conditions of synthesis was determined component by component in the following way. The components of the description corresponding to a serial number of the autoclave, the time of day, and the work-shift number were given random values. The similarities of the values of these components were defined as follows:

$$x\wedge y=\begin{cases}x, & \text{if } x=y;\\ 0, & \text{if } x\neq y.\end{cases}$$

For those components of the description which took numeric values, the similarity was specified after the entire set of possible values from the "similarity interval" was classified by experts. Within such intervals, the values

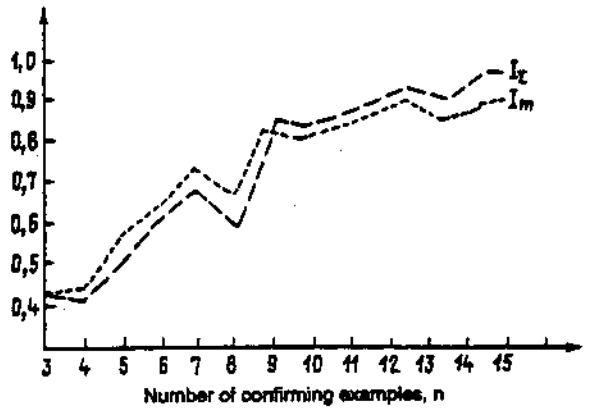


Fig. 2

were considered similar; outside them they were different. More exactly, let L be the set of all similarity intervals for a certain component j that takes numeric values.

In that case,

$$x \wedge y = \begin{cases} l, & x \in l, y \in l, \ l \in L; \\ p, & x \in l, \ y \in k, \ l \in L, \ k \in L, \ l \neq k. \end{cases}$$

In this fashion, we specify similarities of values of the components representing the weight share of volatile substances, acidity, coloration, downtime, and the time from the end of degassing to unloading.

Each of the four polymer properties was investigated separately in the experiment. Each of the 253 situations was classified for each property as belonging to the class of positive or negative examples.

A large number of hypotheses was obtained by the "counterexample prohibition" rule (from 500 to 1500 for the individual polymer properties).

The stability indices I_m, I_Σ were calculated for the hypotheses by a program written by Ivashko for IBM PC AT computer in the C language. The available memory (approximately 217 kB of free working memory) allows the program to compute exact values for hypotheses with $n < 16$. The maximal computation time (for $n = 15$) was 10 sec (Intel 80386 processor, 20 mHz). For hypotheses with large n, upper estimates were calculated in 10 sec according to Theorem 5.5 and 5.6. However, these estimates differed little from 1. Thus, an adequate notion of stability indices could be obtained only for $n < 16$, which was sufficient for this problem: few hypotheses had more than 15 confirming examples.

On this experimental material we confirmed the notion (§4) of the growth of the values of I_Σ with increasing n as a general proposition: in all the four characteristics, the value of I_Σ averaged for all hypotheses either increased monotonically or pased through several small segments of decrease. The value of I_m was slightly greater than I_Σ for small n, and it became smaller for larger n (for the interval (3, 15)). Figure 2 represents a typical behavior of the average values of I_m, I_Σ as a function of n (concerning hypotheses for the causes of absence of non-cut-through areas). For all the eight hypotheses (concerning different properties), identified by experts as the most meaningful ones, the values of integral and average stability indices were above average, which indicates a good selectivity of these indices for this problem.

7. Other Possible Definitions of Stability

The notion of stability can be expressed by various stability indices. In certain circumstances, some may be preferable to others. An analysis of these circumstances is an interesting subject in its own right. Here, we will merely list some possible other definitions of the indices of stability of a hypothesis $<H, \{X_1, ..., X_n\}>$:

1) "Middle layer"

$$I_M = \begin{cases} I_{\frac{n}{2}}, & \text{if } n \text{ is even;} \\ \dfrac{I_{\left[\frac{n}{2}\right]} + I_{\left[\frac{n}{2}+1\right]}}{2}, & \text{if } n \text{ is odd;} \end{cases}$$

2) "Minimal covering"

$$I_c = I_j, \text{ where } j = \min_{2<i<n-1} (I_i \neq 0);$$

3) "Maximal anticovering"

$$I_s = I_j, \text{ where } j = \max_{2<i<n-1} (I_i \neq 1);$$

4) Stability of prediction along the lines of reasoning in [17].

Let S^+ and S^- be the sets of initial examples, $S^q \subset S^T$ is the set of issues (i.e., objects from S^T, for which the prediction must be made), P^+ and P^- are the sets of all (+)- and (-)-predictions obtained on the basis of all hypotheses generated from S^+ and S^-. In that case, I_u is the share of all subsets of the set $S^+ \bigcup S^-$ for which the sets of all predictions generated coincide with predictions obtained from the entire set of examples $S^+ \bigcup S^-$.

The author thanks V. K. Finn, M. V. Arapov, and O. G. Gorbachev for noting the connection between the constructs introduced here and the ideas of nonparametric statistics and providing a brief proof of Lemma 3.1, D. P. Skvortsov for pointing out some inaccuracies in the original draft, and V. G. Ivashko for writing the program of exact calculation of I_Σ and I_m.

REFERENCES

1. D. Reisin, editor, Classification and Clusters [Russian translation], Mir, Moscow, 1980.

2. Yu. A. Shreider, Equality, Similarity, Order [in Russian], Nauka, Moscow, 1971.

3. Y. Kodratoff and G. Tecucci, "Learning based on conceptual distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 10, no. 6, pp. 897-909, 1988.

4. K.-S. Fu, "A graph distance measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. SMC, no. 14, pp. 398-408, 1984.

5. S. A. Vere, "Inductive learning of relational productions," in: Pattern-Directed Inference Systems, eds. D. A. Waterman and F. Hayes-Roth, Academic Press, New York, 1978.

6. V. K. Finn, "Plausible inferences and plausible reasoning," Itogi Nauki i Tekhniki, Ser. Teoriya Veroytnostei, Matematicheskaya Statistika, Teoreticheskaya Kibernetika, vol. 28, pp. 3-84, 1988.

7. S. M. Gusakova and V. K. Finn, "Formalization of local and global similarity," Nauchno-Tekhnicheskaya Informatsiya, Ser. 2, no. 6, pp. 16-19, 1986.

8. S. M. Gusakova and V. K. Finn, "Similarity and plausible inferences," Izv. Akad. Nauk SSSR, Ser. Tekhnicheskaya Kibernetika, no. 5, pp. 42-63, 1987.

9. F. A. Akinniyi and A. K. C. Wong, "A new product graph based algorithm for subgraph isomorphism," in: Proceedings of Pattern Recognition and Computer Vision (June 1983), pp. 457-467, 1983.

10. N. G. Zagoruiko, V. A. Skorobogatov, and P. V. Khvorostov, "Analysis and recognition of molecular structure on the basis of common fragments," Vychislitel'nye Sistemy, iss. 103, pp. 26-50, 1984.

11. E. Yu. Denishchik, "Finding maximal common subgraphs in a family of graphs," Vychislitel'nye Sistemy, iss. 103, pp. 85-89, 1987.

12. S. O. Kuznetsov, "Defining similarity on hypergraphs as the basis for plausible inferences on structured data," in: Proceedings of a National Conference on Artificial Intelligence, Vol. 1 [in Russian], VINITI, pp. 442-448, 1988.

13. S. O. Kuznetsov and V. K. Finn, "Extension of JSM-like expert system procedures onto graphs," Izv. Akad. Nauk SSSR, Ser. Tekhnicheskaya Kibernetika, no. 5, pp. 4-11, 1988.

14. S. O. Kuznetsov, "JSM method as a machine learning system," Itogi Nauki i Tekhniki, Ser. Informatika, Intellektual'nye Informatsionnye Sistemy, vol. 15, 1991 (in press).

15. B. P. Kononov, "Similarity and difference: Operational modeling and aspects of interrelationship and uniqueness," Nauchno-Tekhnicheskaya Informatsiya, Ser. 2, no. 3, pp. 27-31, 1983.

16. B. Efron, Unconventional Methods of Mutlidimensional Statistical Analysis [Russian translation], Finansy i Statistika, Moscow, 1988.

17. V. A. Gusev, "Subsamples and stability concepts used in determinations of the general form of a relation sought for," Zavodskaya Laboratoiya, vol. 53, no. 1, 1987.

18. E. Dantzin, "Algorithms for probabilistic inference," Lecture Notes in Computer Science, vol. 417, 1990.

19. R. Carnap, The Logical Foundations of Probability, University of Chicago Press, Chicago, 1962.

20. L. G. Valiant, "The complexity of computing the permanent," Theoretical Computer Science, no. 8, pp. 189-201, 1979.

21. L. G. Valiant, "The complexity of enumeration and reliability problems," SIAM J. Comput., vol. 8, no. 1, pp. 410-421, 1979.

27 November 1990

THE ALLERTON PRESS JOURNAL PROGRAM

AUTOMATIC DOCUMENTATION & MATHEMATICAL LINGUISTICS

Selected major articles from

NAUCHNO-TEKHNICHESKAYA INFORMATSIYA

Seriya 2. Informatsionnye Protsessy i Sistemy

ALLERTON PRESS, INC.
150 Fifth Avenue New York, N.Y. 10011

ISSN 0005-1055

AUTOMATIC DOCUMENTATION AND MATHEMATICAL LINGUISTICS

(Nauchno-Tekhnicheskaya Informatsiya, Seriya 2)

Vol. 24, No. 6

ALLERTON PRESS, INC.