DISTANCES BETWEEN PREDICATES IN BY-ANALOGY REASONING SYSTEMS

V. Koval, Yu. Kuk

Abstract: *The purpose is to develop expert systems where by-analogy reasoning is used. Knowledge "closeness" problems are known to frequently emerge in such systems if knowledge is represented by different production rules. To determine a degree of closeness for production rules a distance between predicates is introduced. Different types of distances between two predicate value distribution functions are considered when predicates are "true". Asymptotic features and interrelations of distances are studied. Predicate value distribution functions are found by empirical distribution functions, and a procedure is proposed for this purpose. An adequacy of obtained distribution functions is tested on the basis of the statistical χ^2–criterion and a testing mechanism is discussed. A theorem, by which a simple procedure of measurement of Euclidean distances between distribution function parameters is substituted for a predicate closeness determination one, is proved for parametric distribution function families. The proposed distance measurement apparatus may be applied in expert systems when reasoning is created by analogy.*

Keywords: *expert systems, production rules, predicates, distances between predicates, by-analogy reasoning.*

Introduction

Partnership systems are known to be the ones [1] able not only to use experts' knowledge, but also to derive themselves new knowledge from data accumulated in memory. They have means used to derive knowledge from data represented as statistical or empirical "object-feature-time"-type tables [2]. While inferences are obtained in traditional expert systems only deductively, partnership systems use additionally inductive inference features, by-analogy reasoning construction facilities and non-monotone reasonings [1]. The by-analogy reasoning creation basis is the rule that resembling conditions entail resembling effects in immediate proximity to known productions. Therefore, to construct a by-analogy reasoning mechanism, one should be able to compare a condition and an effect resemblance degree. A knowledge in expert systems is usually represented as "if $X_1 \,\&\, X_2 \,\&\ldots\&\, X_m$, then A"–type productions. Compare two productions, for instance, by some PROLOG language features, and left and right sides of both productions are compared. Productions coincide if compared predicates fully coincide. If productions do not coincide, then partnership systems take a non-coincidence degree into account. For this purpose, a distance between predicates is introduced in such systems, and it becomes possible to measure a degree to which one production resembles another. Thus, it is also possible to construct a by-analogy reasoning inference mechanism. By-analogy reasonings may be illustrated by the following example. Assume that it is necessary to check whether conditions $X_1 \,\&\, X_2 \,\&\ldots\&\, X_m$ lead to an effect A. An inference system detects that a knowledge base (KB) contains a resembling knowledge, i.e. "if $Y_1 \,\&\, Y_2 \,\&\, X_3 \,\&\ldots\&\, X_m$, then A ", a truth of which is equal to P. The conditions Y_1 and Y_2 do not coincide with X_1 and X_2 in this knowledge. Their non-coincidence degree is calculated. Hence, find a distance $d(X,Y)$ between the predicates $X = X_1 \,\&\, X_2$ and $Y = Y_1 \,\&\, Y_2$, and, if it does not exceed a threshold η, the conclusion is that A is probable. The truth of this inference is $P' < P$. A truth lowering value depends on a length of a distance between X and Y. The by–analogy inference rule scheme may be represented as

$$\frac{B', B \to A, d(B', B) < \eta}{A}. \tag{1}$$

Example 1. Let a predicate subject domain be a set of real functions $f(x)$ with one variable. Consider three predicates: 1) predicate B', i.e. "to be function $\frac{\sin(x)}{x}$"; 2) predicate B, i.e. "to be polynomial $f_n(x)$ with power exponent *n=2m*", where

$$f_n(x) = 1 - \frac{x^2}{2\cdot 3} + \frac{x^4}{2\cdot 3\cdot 4\cdot 5} - \frac{x^6}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7} + \cdots + (-1)^m \frac{x^{2m}}{2\cdot 3\cdot 4\cdot\cdots\cdot(2m+1)} \text{»;} \qquad (2)$$

and 3) predicate A, i.e. "to be represented as product of linear co-factors $f(x) = \prod_{i=1}^{n}(x-\alpha_i)$, where $\alpha_i, i = 1,\ldots,n$ are roots of equation $f(x) = 0$". The expression $B \to A$ is known [3]. Calculate the distance between B and B' by the following formula: $d(B',B) = \sup_x |g(x) - f_n(x)|$.

When n is chosen, this distance can be made shorter than any number η that is as low as possible: $d(B',B) \le \eta$. This fact be proved, if the function $g(x) = \frac{\sin(x)}{x}$ is expanded into Taylor series:

$$g(x) = \frac{\sin(x)}{x} = 1 - \frac{x^2}{2\cdot 3} + \frac{x^4}{2\cdot 3\cdot 4\cdot 5} - \frac{x^6}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7} + \cdots + (-1)^m \frac{x^{2m}}{2\cdot 3\cdot 4\cdot\cdots\cdot(2m+1)} + \ldots. \qquad (3)$$

Since $d(B',B) \le \eta$, then $B' \to A$ is the by-analogy inference (expression (1)), i.e. the function $g(x) = \frac{\sin(x)}{x}$ can also be expanded into linear co-factors. Since the roots of the equation $g(x) = \frac{\sin(x)}{x} = 0$ are $\pi, -\pi, 2\pi, -2\pi, \ldots$, then, when obtained by analogy, expansion (3) has the following form: $\frac{\sin x}{x} = (1 - \frac{x^2}{\pi^2})(1 - \frac{x^2}{4\pi^2})\cdots(1 - \frac{x^2}{n^2\pi^2})\cdots$. Pursuant to this formula, it is possible to determine the factor under x^2, i.e. $-(\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \cdots)$, and to make the latter equal to the factor under x^2, i.e. to $-\frac{1}{2\cdot 3}$, in expansion (3). And $\frac{1}{2\cdot 3} = \frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \cdots$ is the result, from which the famous Euler formula follows:

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} \cdots = \frac{\pi^2}{6}.$$

1. Distances between Predicates

1.1 Empirical Predicate Distribution Functions. An m-ary predicate $X = X(y_1,\ldots,y_m)$ is understood as a function, values of which are statements about *m* objects. Such objects are predicate argument values. A predicate is an object "feature" under $m = 1$ and it is a "relation" between *m* objects under $m > 1$.

Introduce the notions of empirical frequencies and of predicate value distribution functions needed in order to compare two "resembling" predicates X and Y. Consider the following cases.

1) $m = 1$ and a number of different true statements about an object feature is finite and equal to K. Bring an integer number, respectively, $1,2,\ldots K$ in correspondence with each such statement. Let there be n objects from some subject domain and, respectively, n true statements about a single feature of every such object. Define an empirical frequency for an i-th statement as $p_i = \frac{k_i}{n}$, where k_i is a number of i-th statements from among a whole number of n true statements. Pursuant to these frequencies, define an empirical distribution function $F_n^*(x)$ as a step function of a real variable x. This function is equal to zero under $x \le 1$, to p_1 under $1 \le x < 2$, to $p_1 + p_2$ under $2 \le x < 3$, … , and is to 1 under $x \ge K$. The derived empirical frequencies p_i, $i = 1,\ldots,K$, $\sum_{i=1}^{n} p_i = 1$, and $F_n^*(x)$ characterize this predicate well enough.

2) And now here is the general case when $m > 1$ and a number of different true statements has a power of a continuum. Construct an empirical distribution function $F_n^*(x)$. Let there be n selections that have *m* objects

from some subject domain and n true statements about relations between m objects from each selection. Bring a real number from the space R_1 in correspondence with each such statement. The result is that there are n numbers $x'_1,\ldots,x'_n$ on the straight line R_1. Arrange these numbers in the ascending order, i.e. the variational series $x_{(1)} \le \ldots \le x_{(n)}$ is formed. Define $F_n^*(x)$ as a step function with the steps equal to $1/n$. It is the function of a real variable x, and it is equal to zero under $x \le x_{(1)}$, to k/n under $x_{(k)} \le x < x_{(k+1)}$, $k = 1,\ldots,n-1$, and to 1 under $x \ge x_{(n)}$.

3) Consider the formula $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_K$. Construct an empirical distribution function $F_n^*(x)$ for the formula X. Let a predicate X_i be m_i-ary, $m_i \ge 1$ and a number of different true statements about m_i objects has a power of a continuum. Then, bring the values of X_i in correspondence with the i-coordinates of the points from the K-dimensional space R_K. Let there be n selections that have m objects ($m = \sum_{i=1}^{u} m_i$) from some subject domain and n true statements about relations between m objects from each selection. To reflect relations between m objects, bring the K-dimensional vector from the space R_K in correspondence with each such statement. The result is that there are n vectors $x'_1,\ldots,x'_n$ from R_K, where $x'_i = (x'_{i,1},\ldots,x'_{i,K})$. Define $F_n^*(x)$, where $x = (x_1,\ldots,x_K) \in R_K$, as follows. Consider a set $B_x = \{y \in R_K : y_i < x_i,\ i = 1,\ldots,K\}$. Denote a number of $x'_1,\ldots,x'_n$ by $\nu(B_x)$ as for the vectors that got into B_x. Assume the following: $F_n^*(x) = \nu(B_x)/n$, $x \in R_K$.

It can be shown by analogy with Glivenko-Cantelli theorem [4] that the following assertion is valid for empirical distribution functions: $F_n^*(x)$ converges under $n \to \infty$ to some single limited predicate distribution function $G(x)$.

1.2. Calculating a Distance between Predicates. Differences in functions of distribution of two predicates or formulas can be used in by-analogy reasoning systems in order to compare two “resembling” predicates or formulas. Let $G(x)$ and $Q(x)$ be predicate value probability distribution functions or formula value probability distribution functions, respectively, for X, the first predicate or formula, and for Y, the second predicate or formula. In practice, empirical distribution functions or distribution function estimates are used as the former ones. They are selected from appropriate standard parametric distribution function families and tested for adequacy. The distribution function estimate derivation methodology is considered below.

Definition 1. *A distance $d(X,Y)$ between predicates or* formulas *X and Y is a distance $d(G,Q)$ between two value distribution functions $G(x)$ and $Q(x)$ when these predicates or* formulas *are true under their values.*

Consider the distance d between two formulas $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_u$ and $Y = Y_1 \,\&\, Y_2 \,\&\ldots\&\, Y_w$ for the case when a “feature” or a “relation”, described by each separate predicate, are by no means associated with “features” or “relations” described by other predicates. Let $G_{X_1}, G_{X_2},\ldots,G_{X_u}$ and $Q_{Y_1}, Q_{Y_2},\ldots,Q_{Y_w}$ be the distribution functions, respectively, for $X_1, X_2,\ldots,X_u$ and $Y_1, Y_2,\ldots,Y_w$. Then, d between $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_u$ and $Y = Y_1 \,\&\, Y_2 \,\&\ldots\&\, Y_w$ is equal to the distance between two products of the respective distribution functions $G_X = G_{X_1} \cdot G_{X_2} \cdot\ldots\cdot G_{X_u}$ and $Q_Y = Q_{Y_1} \cdot Q_{Y_2} \cdot\ldots\cdot Q_{Y_w}$.

If an expression for d between predicates or formulas is chosen correctly, it is possible to use further on “good” features of this distance, for instance, the distance calculation procedure itself may be simplified. Consider various expressions used to calculate distances between X and Y. The distance

$$d(X,Y) = d(G,Q) = \sup_x \left| G(x) - Q(x) \right| \quad (4)$$

means an absolute deviation of values for one distribution function with respect to another distribution function at each point and the distance

$$d(X,Y) = d(G,Q) = \int (G(x) - Q(x))^2 \, dQ(x) \quad (5)$$

takes a root mean square deviation of these values into account.

Example 2. Calculate the distance d between the predicates B' and B from Example 1. For every x, the real value for B' is equal to $g(x) = \frac{\sin(x)}{x}$. Therefore, the distribution function $G(y)$ for this predicate is equal to zero under $y < g(x)$ and to 1 under $y \geq g(x)$. The values of B correspond to the values of the polynomial $f_n(x)$: $y_1 = f_1(x)$, $y_2 = f_2(x)$, ..., $y_n = f_n(x)$ that, under different and sufficiently large n, $n > n_1$, are arranged in a certain way within the interval Δ of the following form: $\Delta = [g(x) - \eta, g(x) + \eta]$. However, when n increases, the points $f_n(x)$ approach the point $g(x)$ because of $f_n(x) \to g(x)$. The distribution functions $Q(y)$ for these points are not found, since only the upper estimate for d between B' and B is important. The following is made: move each of these points away from $g(x)$ in such a way that they fill in the interval Δ uniformly. The result is that the distribution function $\tilde{Q}(y)$ in its new position becomes uniform, but the distance between $G(y)$ and the new $\tilde{Q}(y)$ increases here in comparison with the previous one between $G(y)$ and $Q(y)$ Therefore: $d(B,B') = d(G,Q) < d(G,\tilde{Q})$. Since

$$\tilde{Q}(y) = \begin{cases} n_1/n, & \text{when } y = g(x) - \eta \\ \frac{n-n_1}{2n\eta} y + \frac{n_1}{n} - \frac{n-n_1}{2n\eta}(g(x) - \eta), & \text{when } g(x) - \eta \leq y \leq g(x) + \eta \\ 1, & \text{when } y \geq g(x) + \eta \end{cases}$$

then, if formula (5) is used, the following expression takes place: $d(G,\tilde{Q}) = \frac{1}{8}(1 - \frac{n_1}{n})^3 \eta < \frac{\eta}{8}$. Hence, the upper estimate is derived for $d(B,B')$. Thus, if $\sup_x |g(x) - f_n(x)| \leq \eta$, then $d(B,B') < \eta/6$. Therefore, these formulas for the distances are equivalent.

1.3 Kulbak–Leibler Distance, χ^2–Distance, Hellinger Distance. Consider now different types of distances between two predicates X and Y for the case when their distribution functions Q and G have, respectively, the densities $q(x)$ and $g(x)$ as for a measure μ. The Lesbegue measure may be used for one group of distribution functions (absolutely continuous distributions) and a counting measure may be taken for another group (discrete distributions) as μ. Let N_Q be a carrier of Q ($N_Q = \{x : q(x) > 0\}$), and let N_G be a carrier of G ($N_G = \{x : g(x) > 0\}$). The Kulbak–Leibler distance between X and Y is calculated in the following way:

$$r_1(X,Y) = r_1(\mathrm{G},Q) = \int_{N_G} \ln \frac{g(x)}{q(x)} g(x) \mu(dx).$$

The χ^2–distance between X and Y is

$$r_2(X,Y) = r_2(\mathrm{G},Q) = \int_{N_Q \cup N_G} \frac{(q(x) - g(x))^2}{g(x)} \mu(dx).$$

The values of $r_1(X,Y)$ and $r_2(X,Y)$ are more than or equal to zero. However, the equalities $r_1(X,Y) = 0$ and $r_2(X,Y) = 0$ are possible only under $Q = G$. Since $r_1(X,Y)$ and $r_2(X,Y)$ are not the symmetric functions of Q and G, then $r_1(X,Y)$ and $r_2(X,Y)$ are not the distances in the general case because of $r_1(X,Y) \neq r_1(Y,X)$ and $r_2(X,Y) \neq r_2(Y,X)$. Nevertheless, essentially speaking and from the statistical point of view, $r_1(X,Y)$ and $r_2(X,Y)$ characterize a deviation of Q from G.

The Hellinger distance between X and Y is

$$r_3(X,Y) = r_3(\mathrm{G},Q) = \int_{N_Q \cup N_G} \left(\sqrt{g(x)} - \sqrt{q(x)}^2 \right) \mu(dx).$$

and it is already the symmetric function for X and Y. The value $\sqrt{r_3(Q,G)}$ possesses all the metric characteristics between the functions $\sqrt{q(x)}$ and $\sqrt{g(x)}$ in the metrical space L_2.

Consider the features of these distances, important when a predicate resemblance threshold is chosen. If a predicate closeness degree is characterized by such distances when $q(x)/g(x)$ is close to 1, then the following result turns out to take place:

$$r_1(Q,G) \approx \frac{1}{2} r_2(Q,G) \approx 2 r_3(Q,G) .$$

Asymptotically, all the distances behave in the same way. To study this asymptotic feature, assume that G and Q for X and Y are taken from one and the same parametric family and defined, respectively, by the parameters θ and $\theta+\Delta$. Then, the rate of the convergence to zero for the distance between X and Y is equal to $O(\Delta^2)$ under $\Delta \to 0$. This fact follows from the asymptotic equality

$$r_3(\Delta) \approx \frac{I(\theta)}{4} \Delta^2 ,$$

where $I(\theta)$ is the Fisher information found by the formula

$$I(\theta) = \int \frac{(g'_\theta(x))^2}{g_\theta(x)} \mu(dx) .$$

1.4 Predicate Comparison Procedure Simplification Theorem. The predicate resemblance determination procedure falls into two stages: 1) calculate a distance between predicates; and 2) compare a calculated distance with a threshold η. Let G and Q be distribution functions for predicates X and Y that belong to the same parametric family $\Psi = (G_\theta | \theta \in \Theta)$ and differ only in their parameters. Assume that G_{θ_1} and G_{θ_2} are, respectively, the predicate value distribution functions for X and Y. Consider the Kulbak-Leibler, χ^2-and Hellinger distances as the ones between predicates: $\rho_i(\theta_1,\theta_2)$, $i=1,2,3$. The following theorem is true.

Theorem 1. *Assume that value distribution functions for predicates X and Y belong to a parametric distribution function family $\Psi = (G_\theta / \theta \in \Theta)$. Let the following conditions be met: 1) a parametric set Θ is compact; 2) $G_{\theta_1} \neq G_{\theta_2}$ under $\theta_1 \neq \theta_2$; 3) for every $\theta \in \Theta$, Fisher information is restricted: $0 < I(\theta) \le 4b < \infty$. Then, $\rho_i(\theta_1,\theta_2) \le \delta$, $i=1,2,3$ is equivalent to $(\theta_1-\theta_2)^2 \le \delta / b_i$, where b_i, $i=1,2,3$ are constant, $b_1 = 2b$, $b_2 = 4b$, $b_3 = b$.*

This theorem reduces the predicate resemblance determination procedure to the simple procedure by which a Euclidean distance between distribution function parameters is determined. The Θ-set compactness condition is not assumed to be restricting and it means that Θ is restricted. The second condition means that $\rho_i(\theta_1,\theta_2) > 0$ takes place under $\theta_1 \neq \theta_2$.

1.5. Distribution Function Estimates. As a rule, a predicate distribution function is not known. It is not very convenient to deal with empirical distribution functions. Therefore, the already known classes of distributions are used and estimates for $G(x)$ are created. Assume that an unknown estimate of $G(x)$ for a predicate X belongs to $\Psi = (G_\theta / \theta \in \Theta)$. Construct an empirical function G_n^* for X. Let $G*$ be a function from Ψ that is closest to G_n^* as for a distance d, i.e.

$$d(G*,G_n^*) = \min_{\Pi\in\Psi} d(\Pi,G_n^*) .$$

$G*$ with the parameter $\theta*$ is an estimate for $G(x)$ as for a minimum of d.

Consider the practical methods used to create the estimates for distribution functions. First of all, describe the χ^2–procedure that helps to find estimates. In this case, the distance

$$d(G,Q)=\sum_{i=1}^{r}\frac{(P_G(\Delta_i)-P_Q(\Delta_i))^2}{P_G(\Delta_i)}$$

is used as d ; $\Delta_1,\ldots,\Delta_r$ are non-intersecting sets of a predicate value space R and their union is equal to R;

$$P_G(\Delta_i)=\int\limits_{\Delta_i}dG(x)\,,\; P_Q(\Delta_i)=\int\limits_{\Delta_i}dQ(x)\,,\; i=1,\ldots,r\,.$$

Take $G_n^*(x)$ as $Q(x)$. The estimate $\theta*$ as for the given minimum distance is a value of θ, and

$$d(G_\theta,G_n^*)=n\sum_{i=1}^{r}\frac{\left(P_\theta(\Delta_i)-\frac{\nu_i}{n}\right)^2}{P_\theta(\Delta_i)}=\sum_{i=1}^{r}\frac{\left(nP_\theta(\Delta_i)-\nu_i\right)^2}{nP_\theta(\Delta_i)}\,. \qquad (6)$$

is minimized under this distance; in the present case, $\nu_i=nG_n^*(\Delta_i)$ is a number of predicate values that got into the set Δ_i and under which a predicate is "true". Differentiate expression (6) with respect to the parameters, the components of which make up the vector θ, make the derivatives equal to zero, and the equation system is derived relative to unknown parameters. Solve this system and find the estimates for the parameters. The obtained $G*(x)$ is then tested for adequacy. If a test result shows that $G*(x)$ is not adequate to the data, then an initial distribution function family should be changed.

Consider the practically important maximum likelihood method also used to derive the estimates. To create a maximally likely estimate means to define one more important distance between an arbitrary Q and G_θ from $\Psi=(G_\theta\,/\,\theta\in\Theta)$. It is assumed that G_θ possesses a density $g_\theta(x)$ with respect to a measure μ. Such a distance is expressed by the formula

$$\rho(G_\theta,Q)=-\int\ln g_\theta(x)Q(dx)\,.$$

If an empirical G_n^* is taken as Q, then the estimate for θ is called the maximum likelihood estimate and it minimizes the distance $\rho(G_\theta,G_n^*)$. The yielded function estimates have the "good" features, i.e. they are efficient and asymptotically not biased.

1.6. Testing for Adequacy. Obtained distribution function estimates are tested for adequacy before a distance between predicates is found by means of them. An adequacy of a found function is tested for by the χ^2–statistics. The testing mechanism is as follows. Consider the hypothesis that, when a predicate is "true", probable predicate values are distributed by $G*(x)$. Divide a predicate value space into a finite number of sets $\Delta_1,\ldots,\Delta_r$ without common points. Calculate the values of $p_i=P_G(\Delta_i)$. Determine the frequencies ν_i, i.e. a number of predicate values under which it is "true" and that got into a set Δ_i. Calculate the statistics

$$\chi^2=\sum_{i=1}^{r}\frac{\left(np_i-\nu_i\right)^2}{np_i}\,.$$

It is possible to show by analogy with [4] that the χ^2-statistics distribution function does not depend on an initial predicate value distribution function at all under $n\to\infty$. The former function is expressed by the formula

$$w_{r-1}(x)=2^{\frac{1-r}{2}}\Gamma^{-1}\left(\frac{r-1}{2}\right)x^{\frac{r-3}{2}}e^{-\frac{x}{2}},\quad x>0$$

and helps to find the point $x_{0.05}$ for which the expression

$$\int\limits_{x_{0.05}}^{\infty}w_{r-1}(x)dx=0.05$$

takes place. If $\chi^2>x_{0.05}$, then the choice of a distribution function is wrong.

2. A By-Analogy Reasoning System Flowchart

Figure 1 depicts a by-analogy reasoning system flowchart. Let the following request be received by the system: “Is effect A possible when conditions $X_1 \& X_2 \& \ldots \& X_m$ are met?” The production rule “if $X_1 \& X_2 \& \ldots \& X_m$ then A” is sought for in the KB. If it is found, the answer is positive. If it is absent, the production rule “if $Y_1 \& X_2 \& \ldots \& X_m$, then A” is sought for, the conditions of which contain the same predicate names as in the request conditions, but the first predicate differs from first predicate in the request. If this rule is not found in the KB, then such a production rule is sought for, the conditions of which differ from the request conditions already in the second predicate: “if $X_1 \& Y_2 \& \ldots \& X_m$ then A”. And so on. For more certainty, let the procedure result be that the desired production rule “if $X_1 \& X_2 \& \ldots$ $\ldots \& X_{i-1} \& Y_i \& X_{i+1} \& \ldots \& X_m$ then A” is found in the KB at the i-th step. However, the predicate Y_i does not coincide in this rule with the predicate X_i. Therefore, the procedure is started up that determines a closeness of predicates that do not coincide.

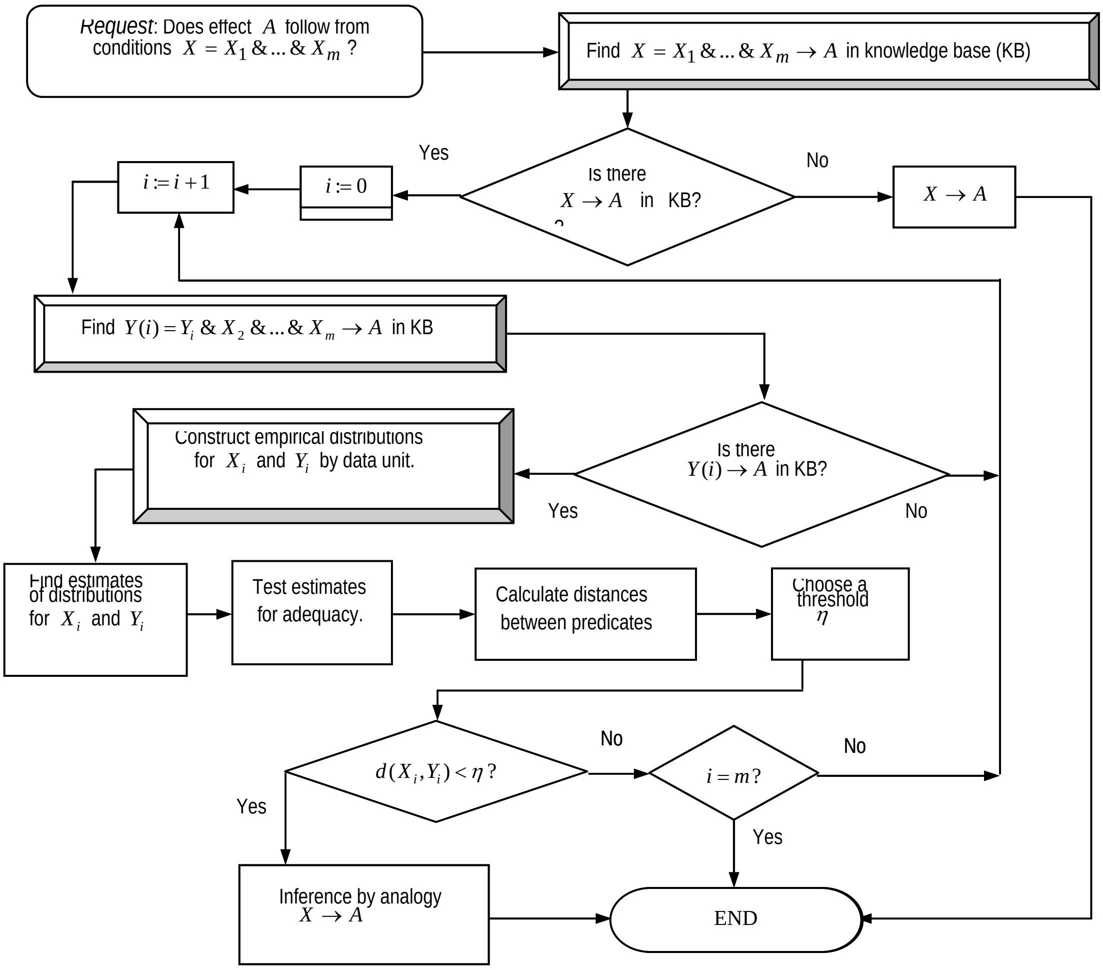
Request: Does effect A follow from conditions $X = X_1 \& ... \& X_m$?
Find $X = X_1 \& ... \& X_m \to A$ in knowledge base (KB)
Yes
No
Is there $X \to A$ in KB?
$i := i + 1$
$i := 0$
$X \to A$
Find $Y(i) = Y_i \& X_2 \& ... \& X_m \to A$ in KB
Construct empirical distributions for X_i and Y_i by data unit.
Is there $Y(i) \to A$ in KB?
Yes
No
Find estimates of distributions for X_i and Y_i
Test estimates for adequacy.
Calculate distances between predicates
Choose a threshold η
$d(X_i, Y_i) < \eta$?
No
$i = m$?
No
Yes
Yes
Inference by analogy $X \to A$
END

Figure 1. By-Analogy Reasoning System: A Flowchart

The data about distribution of the values of X_i and Y_i, under which they are “true”, are extracted from the database. Pursuant to these data, the empirical $F_n^*(x)$ and $G_n^*(x)$ are constructed for the values under which they are “true”. In accordance with $F_n^*(x)$ and $G_n^*(x)$, the estimates of $F^*(x)$ and $G^*(x)$ are found as for X_i and Y_i. To find such estimates, introduce a distance between the distribution functions. To derive the estimates, find the distribution functions from the specified families that are closest to the found empirical functions in the sense of an introduced distance metrics. The obtained estimates for $F^*(x)$ and $G^*(x)$ are then tested for their adequacy as for the available empirical data by the χ^2–criterion. If the estimates for $F^*(x)$ and $G^*(x)$ do not fit available empirical data, choose another family where the same estimates are sought for again. The adequate estimates of $F^*(x)$ and $G^*(x)$ are yielded, and a distance d between the considered predicates is calculated by means of them. This distance determines a degree of "resemblance" or "closeness" for X_i and Y_i. Predicates are close if a distance between them does not exceed some threshold. As a threshold, a sufficiently small positive number η is chosen, and a value of this number states a by-analogy inference truth. Under $d(X_i, Y_i) \le \eta$, there is the following by-analogy inference: "if $X_1 \,\&\, X_2 \,\&\ldots\&\, X_m$ then A". If $d(X_i, Y_i) > \eta$ takes place, then a found production rule is rejected, and a new production rule is sought for that differs from a required one in a next-coming $(i+1)$ -th predicate.

Conclusion

The paper considers different-type distances between predicates. They are the distances between predicate value distribution functions under which predicates are "true". The asymptotic features of such distances and the interrelation between the latter are studied. The paper proposes the procedure used to find distributions of predicate values for the case when predicates are true. The distribution functions are found by the empirical distribution ones. The paper also deals with the mechanism that tests an adequacy of a yielded distribution function on the basis of the χ^2–criterion. The predicate resemblance determination procedure is replaced by the simple procedure that determines Euclidean distances between distribution function parameters. The replacement theorem is proved for the parametric families. The proposed distances can be used in expert systems in order to construct by-analogy reasonings.

Bibliography

[1] N.G. Zagoruyko. Application Methods for Data and Knowledge Analysis. Novosibirsk, 1999, 269p.

[2] V.N. Koval, Yu.V. Kuk. Finding Unknown Rules of an Environment by Intelligent Goal-Oriented Systems, "Information Theories and Applications", International Journal, vol. 17, N 3, p. 127-138, Sofia, 2001.

[3] G. Polya. Mathematics and Plausible Reasoning. Prinston, New Jersey. 1954, 464p.

[4] A.A. Borovkov. Mathematical Statistics. Moscow, Nauka, 1984, 472p.

Author information

Valeriy Koval – Institute of Cybernetics, Head of Department, address: 03680, Kiev, Prospect Glushkova, 40, Ukraine; e-mail: icdepval@ln.ua

Yuriy Kuk - Institute of Cybernetics, senior scientific researcher, Ukraine; Kiev, e-mail: vkyk@svitonline.com .