OPTIMAL QUESTIONNAIRES WITH UNEQUAL COSTS OF QUESTIONS

1969

SovietRxiv

Abstract

Full Text

UDC 519.95

CYBERNETICS AND CONTROL THEORY

P. P. PARKHOMENKO

OPTIMAL QUESTIONNAIRES WITH UNEQUAL COSTS OF QUESTIONS

(Presented by Academician V. A. Trapeznikov on 6 V 1968)

There is a finite set E of events $y_i,\ i = 1, 2, \ldots, N$. To each event $y \in E$ there is assigned a (relative) weight $p(y)$. We restrict ourselves to the case where $0 < p(y) < 1$ and

$$\sum_{y \in E} p(y) = 1\ ^*.$$

Also given is a set T_1 of partitions $t_j,\ j = 1, 2, \ldots, |T_1|$, of the set E into classes. The elements $t \in T_1$ will be called questions. The number $a(t)$, $2 \leq a(t) \leq N$, of classes $E_{\mu(t)},\ \mu(t) = 1, 2, \ldots, a(t)$, in the partition $t \in T_1$ is called the arity of the question t; the classes $E_{\mu(t)}$ will be called the outcomes, or answers, of the question t. To each question t a cost $c(t) > 0$ is assigned.

Form the system of sets whose elements $\mathcal{E}$ are the set E, the classes $E_{\mu(t_j)}$, and all possible nonempty intersections

$$E_{\mu(t_{j_1})} \cap \ldots \cap E_{\mu(t_{j_i})} \cap \ldots \cap E_{\mu(t_{j_u})},$$

$$t_j \in T_1, \quad \mu(t_{j_i}) = 1, 2, \ldots, a(t_{j_i}), \quad u = 2, 3, \ldots, |T_1|.$$

By the symbol τ_j we shall denote the partition of the set $\mathcal{E}$ into classes $\mathcal{E}_{\mu(\tau_j)} = E_{\mu(t_j)} \cap \mathcal{E},\ \mu(\tau_j) = 1, 2, \ldots, a(\tau_j)$. Obviously, $a(\tau_j) \leq (t_j)$. Let us set $c(\tau_j) = c(t_j)$. All questions τ with arity $a(\tau) \geq 2$ form a set T. For $\mathcal{E} = E$ we have $\mathcal{E}_{\mu(\tau^j)} = E_{\mu(t^j)}$, i.e. $T_1 \subseteq T$.

Consider a directed graph $G = (Z, \Gamma)$ with root x_0 and such that $Z = \{z \mid z \in Q \cup K\}$, where $Q \cap K = \emptyset$; $z = x \in Q \Rightarrow 2 \leq |\Gamma x| \leq N$; $z = k \in K \Rightarrow |\Gamma k| = 0$; $z \in Z \setminus x_0 \Rightarrow |\Gamma^{-1} z| > 0$; $|\Gamma^{-1} x_0| = 0$; the root $x_0 \in Q$ is connected with every other vertex z by at least one path. Assign $x = \tau \in T$, with $|\Gamma x| = a(x) = a(\tau)$. Then the arcs $(x, \Gamma x)$ correspond to the pairs $(\tau, \mathcal{E}_{\mu(\tau)})$, and the paths from x_0 to $k \in K$ correspond to sequences of pairs

$$\alpha_k = \langle(\tau_1, \mathcal{E}_{\mu(\tau_1)}), \dots, (\tau_r, \mathcal{E}_{\mu(\tau_r)}), \dots, (\tau_{\nu_k}, \mathcal{E}_{\mu(\tau_{\nu_k})})\rangle,$$

and, finally, to the vertices $k \in K$ correspond the outcomes $\mathcal{E}_{\mu(\tau_{\nu_k})} = E_k \subseteq E$ of the questions τ_{ν_k}. If in G, for every path from x_0 to $k \in K$, we have $\tau_1 = t \in T_1$ and $\mathcal{E}_{\mu(\tau_r)} = E_{\mu(t_r)} \cap \mathcal{E}_{\mu(\tau_{r-1})}$, $r = 1, 2, \dots, \nu_k$, $\mathcal{E}_{\mu(\tau_0)} = E$, and also, in addition, to the root and internal vertices $x \in Q$ there are assigned costs of questions $c(x) = c(\tau)$, and to terminal vertices $k \in K$ weights

$$p(k) = \sum_{y \in E_k} p(y),$$

with

$$\bigcup_{k \in K} E_k = E,$$

then, following (1), we shall call the graph G a questionnaire for E. In (1) questionnaires for which $c(\tau) = 1$ for all $\tau \in T$ are studied; here this restriction is removed.

We shall say that a subset E_k of events $y \in E$ is identified by the sequence of questions

$$\beta_k = \langle\tau_1, \dots, \tau_r, \dots, \tau_\nu\rangle,$$

if

---

* In the general case, if $w(y) > 0$ is the complete weight of the event $y \in E$, then $p(y) = w(y)/W$, where $W = \sum_{y \in E} w(y)$.

$$\bigcap_{r=1}^{k} \mathcal{E}_{\mu(r_p)} = E_k.$$

We shall call the sequence β_k irredundant if not a single question can be removed from it without changing the subsets E_k identified by this sequence. Below we consider questionnaires for which $K = E$, $E_k = y_i \in E$, $i = 1, 2, \dots, N$, and each event y is identified by one irredundant sequence of questions; such questionnaires can be represented by a tree (1), a pragraph (2), with root x_0.

Let the questionnaire G contain q_m questions with base a_m, $m \in M$, and q_l questions with price c_l, $l \in L$ (M, L are certain numerical sets). Then

$$|Q| = \sum_{m \in M} q_m = \sum_{l \in L} q_l$$

and

$$N = \sum_{m \in M} q_m (a_m - 1) + 1. \tag{1}$$

With respect to a vertex z, we shall distinguish the sets Γz–its successors, $\hat{\Gamma} z \setminus z = \Gamma z \cup \Gamma(\Gamma z) \cup \Gamma(\Gamma(\Gamma z)) \cup ...$–its descendants, $\Gamma^{-1} z$–its predecessors, and $\hat{\Gamma}^{-1} z \setminus z = \Gamma^{-1} z \cup \Gamma^{-1}(\Gamma^{-1} z) \cup ...$–its ancestors. The path from x_0 to z will be denoted by $s[x_0, z]$. The price of the path $s[x_0, z]$ is the sum

$$c(x_0, z) = \sum_{x \in \hat{\Gamma}^{-1} z / z} c(x). \tag{2}$$

Set $c(x_0, x_0) = 0$. The price $c(x_0, y)$ of the path $s[x_0, y]$ characterizes the expenditures for identifying the event y. We shall call a traversal of the graph G the set of paths $\{s[x_0, y_j] \mid y_j \in E\}$. Then the quantity

$$C(x_0, E) = \sum_{y_i \in E} c(x_0, y_i) p(y_i), \tag{3}$$

called the price of the questionnaire traversal, characterizes the mean weighted expenditures for identifying events over the questionnaire as a whole. The traversal price is a sufficiently general and at the same time practically useful characteristic of a questionnaire. Thus, for questionnaires with equal question prices, from (3) one obtains the traversal length of a questionnaire introduced in (1), corresponding, for example, to the mean length of a code combination in Shannon–Fano codes (3) or to the mean number of operations in sorting problems (4). If the price of a question is the cost or time of carrying out an individual check, and the weight of an event is the probability of the (serviceable or faulty) state of the object under inspection, then (3) gives the mean cost or mean time for determining one state in a conditional diagnostic procedure.

A questionnaire G for E having the minimum traversal price will be called optimal and denoted by G_0. The note is devoted to studying the properties of optimal questionnaires. The basis of the study is transformations of the questionnaire G that are invariant with respect to q_m and q_l.

The weight of a question $x \in Q$ is the quantity

$$p(x) = \sum_{y \in \hat{\Gamma} x \cap E} p(y). \tag{4}$$

The root G_0 has the greatest weight.

From (2), (3), and (4) we obtain:

$$C(x_0, E) = \sum_{x_i \in Q} c(x_i)p(x_i). \tag{5}$$

Let $G(Z, \Gamma)$ be a questionnaire and let z_1 be its vertex. The subquestionnaire G_1 with root z_1 of the questionnaire G is the graph $G_1 = (Z_1, \Gamma_1)$, where $Z_1 = \hat{\Gamma} z_1 \subseteq Z$

and $\Gamma_1 z = \Gamma z \cap Z_1$. The subquestionnaire G_1 is a questionnaire for $E_1 = \hat{\Gamma} z_1 \cap E$; if $z_1 \in E$, then G_1 is a degenerate subquestionnaire.

I. Let us single out in G two subquestionnaires G_1 and G_2 with roots z_1 and z_2, respectively, where $\hat{\Gamma} z_1 \cap \hat{\Gamma} z_2 \emptyset$. The operation of interchanging the subquestionnaire G_1 with the subquestionnaire G_2 defines a new questionnaire G^T, obtained from G by interchanging the endpoints (or origins) of the arcs $(\Gamma^{-1} z_1, z_1)$ and $(\Gamma^{-1} z_2, z_2)$, i.e. $z_1^T = z_2$ and $z_2^T = z_1$, with subsequent recalculation of the weights of the vertices. The difference of the traversal costs $C(x_0, E)$ of the questionnaire G and $C(x_0^T, E)$ of the questionnaire G^T is equal to

$$\Delta^T C = [c(x_0, z_1) - c(x_0, z_2)][p(z_1) - p(z_2)]. \tag{6}$$

Lemma 1. To the vertices G_0, arranged in nondecreasing order with respect to the costs of the paths leading to them, weights are assigned in nonincreasing order.

II. Let us single out in G two questions x_1 and x_2 with costs $c(x_1)$ and $c(x_2)$, respectively. The operation of interchanging the cost of the question $c(x_1)$ with the cost of the question $c(x_2)$ defines a new questionnaire G^c, obtained from G by assigning to the question x_1 the cost $c(x_2)$ and to the question x_2 the cost $c(x_1)$, followed by recalculation of the costs of the paths. The difference of the traversal costs of the questionnaires G and G^c is equal to:

$$\Delta^c C = [c(x_1) - c(x_2)][p(x_1) - p(x_2)]. \tag{7}$$

Lemma 2. To the questions G_0, arranged in nondecreasing order with respect to their costs, weights are assigned in nonincreasing order.

III. From Lemmas 1 and 2 it follows:

Lemma 3. To the vertices G_0, arranged in nondecreasing order with respect to their ranks, weights are assigned in nonincreasing order and the costs of the paths leading to them are in nondecreasing order; moreover, to the questions

are assigned their costs in nondecreasing order. Vertices of different ranks have different costs of the paths leading to them.

IV. Let us single out in G two questions x_1 and x_2 according to their bases $a(x_1)$ and $a(x_2)$, respectively, where $a(x_1)-a(x_2)=n>0$. We artificially increase the base $a(x_2)$ to the value $a(x_1)$ by adding, as successors of the question x_2, n fictitious events $x_{2,j}$, $j=1,2,\dots,n$, with zero weights*. The operation of interchanging the bases $a(x_1)$ and $a(x_2)$ defines a new questionnaire G^a, obtained from G by interchanging n degenerate subquestionnaires whose roots are the fictitious vertices $x_{2,j}$, with n subquestionnaires with roots $x_{1,i_\alpha}\in\{x_{1,i}\mid i=1,2,\dots,a(x_1)\}=\Gamma x_1$, followed by deletion of the n successors of the question x_1^a with zero weights and recalculation of the vertex weights, as well as the costs of the paths. Then, according to (6), we obtain:

$$\Delta^a C=[c(x_0,x_{1,i})-c(x_0,x_{2,j})]\sum_{\alpha=1}^{n}p(x_{1,i_\alpha}). \tag{8}$$

Lemma 4. To the successors of the questions G_0, arranged in nonincreasing order with respect to their bases, are assigned the costs of the paths leading to them in nondecreasing order.

Corollary. To the questions G_0, arranged in nonincreasing order with respect to their bases, are assigned the costs of the questions and the costs of the paths leading to them in nondecreasing order. In this case the questions are arranged in nonincreasing order with respect to their ranks and in nonincreasing order with respect to their weights.

V. Let us single out in G, satisfying the conditions of Lemmas 1-4 and the corollary, two questions x_1 and x_2 and their successors $x_{1,i}\in\Gamma x_1$ and $x_{2,j}\in\Gamma x_2$. Let $a(x_1)\geq a(x_2)$. Construct the questionnaire G' by redistributing the weights

* This, obviously, will not change the traversal cost of G.

$p(x_{1,i})$ and $p(x_{2,j})$ in such a way that the $a(x_1)$ largest weights are assigned to $x_{1,i}\in\Gamma x_1'$ and the $a(x_2)$ smallest ones to $x_{2,j}\in\Gamma x_2'$. In this case $p(x_1')=p(x_1)+\delta$ and $p(x_2')=p(x_2)-\delta$, $\delta\geqslant 0$. Analysis of the results of permuting the pairs of subquestionnaires with roots x_1' and x_2', and also x_M' and x_2', where $x_M'\in\Gamma x_1'$, with

$$p(x_M')=\max_{i,j}[p(x_{1,i}),\,p(x_{2,j})],$$

and of permuting the costs of the questions of the indicated pairs of roots (when $x_M'\in Q$) gives:

Lemma 5. In G_0, among vertices with identical costs of the paths leading to them, there is not a single one whose weight exceeds the sum of the weights of the

other a vertices with the same cost of the paths leading to them, where a is the smallest of the bases of the questions of which the vertices under consideration are successors.

VI. **Theorem.** An optimal questionnaire G_0, allowing one to distinguish N events by means of

$$\sum_{m \in M} q_m = \sum_{l \in L} q_l$$

questions having bases a_m and costs c_l, is a pruned tree with root x_0 in which

$$N = \sum_{m \in M} q_m (a_m - 1) + 1,$$

and such that its vertices, arranged in nondecreasing order with respect to their ranks, are assigned weights in nondecreasing order, and the costs of the paths leading to them are in nonincreasing order; moreover, the questions are arranged in nondecreasing order with respect to their bases and in nonincreasing order with respect to their costs; among all vertices having the same cost of the paths leading to them as the successors of a question with base a, there is not a single one whose weight exceeds the sum of the weights of the other a vertices with the same cost of the paths leading to them; the search cost is equal to the sum of the products of the costs of the questions by their weights, or to the sum of the products of the costs of the paths leading to the events by the weights of the events.

Institute of Automation and Telemechanics
(Technical Cybernetics)

Received
22 IV 1968

REFERENCES

1. C. Picard, *Théorie des questionnaires*, Paris, 1965.
2. K. Berge, *Theory of Graphs and Its Applications*, Moscow, IL, 1962.
3. B. E. Shannon, Works on Information Theory and Cybernetics, Moscow, IL, 1963, p. 270.
4. W. H. Burge, Sorting, Trees and Measures of Order, *Information and Control*, 1, No. 3, 181 (1958).

Note: Figure translations are in progress. See original paper for figures.