KYBERNETIKA — VOLUME *28* (1992), NUMBER 5, PAGES 392 – 401

ON VARIOUS CRITERIA OF OPTIMALITY IN PROBABILISTIC DECISION–MAKING

Jiřina Vejnarová

There are two possibilities how to define the optimality of a decision function – with respect to a given set of distributions – in "local" and "global" senses, applying the minimax rule. But an optimal decision function in any of these senses need not be the optimal one for any distribution in this set. It is shown, using linear programming methods, how to find out whether the global minimax decision function is optimal or not. A suitable representation of the decision function is found – in the latter case – on the base of a barycenter concept.

1. INTRODUCTION

Let us consider the following problem. We have to determine a value of a variable Y knowing the value of a variable X. A *decision function* is a mapping

$$d : \mathbf{X} \longrightarrow \mathbf{Y}$$

(where $\mathbf{X}$ and $\mathbf{Y}$ are the ranges of X and Y respectively), which assigns a value y to every x. Let us suppose P to be a joint distribution of XY. Then we can define an *error of the decision* $d(x)$ as

$$e_P(d(x); x) = \sum_{y \in \mathbf{Y} : y \neq d(x)} P(x, y),$$

and an *expected error* of the decision function d as

$$\bar{e}_P(d) = \sum_{x \in \mathbf{X}} e_P(d(x); x) = \sum_{(x,y) \in \mathbf{X} \times \mathbf{Y}} P(x, y)(1 - \delta(d(x), y))$$

(where $\delta(u, v) = 1$ if $u = v$ and $\delta(u, v) = 0$ otherwise).

Our aim is to provide such decisions, which minimize the decision error.

If we knew the distribution P, we could choose a decision function d_P satisfying inequalities

$$e_P(d_P(x); x) \leq e_P(d(x); x)$$

for every $x \in \mathbf{X}$ and every decision function d, or equivalently

$$\bar{e}_P(d_P) \leq \bar{e}_P(d)$$

for every decision function d. Any function d_P satisfying these inequalities is called an *optimal decision function* with respect to the distribution P.

But in many cases we do not know the distribution P exactly. Let us consider the situation in which we can assume that it belongs to some class of distributions $\mathcal{P}$. If we make a decision $d(x) = y$, the value of the corresponding error can be as large as

$$\max_{P\in\mathcal{P}} \; e_P(y;x).$$

If we wish this maximum error to be the least possible one, we have to choose such y that minimizes this expression. Then, we get a *minimax rule*

$$\min_{y} \; \max_{P\in\mathcal{P}} \; e_P(y;x).$$

Any optimal decision function (it need not be unique) with respect to this rule (assuming $\mathcal{P}$ is fixed) will be denoted d_*:

$$d_*(x) \in \arg\min_{y} \; \max_{P\in\mathcal{P}} \; e_P(y;x)$$

and called a *local minimax decision function.*

But there is another possibility of defining the minimax decision function. We can select a function d which minimizes the maximal possible expected error of the decision function, i.e.

$$\max_{P\in\mathcal{P}} \; \bar{e}_P(d).$$

Any such function is called a *global minimax decision function* and denoted d^*:

$$d^* \in \arg\min_{d} \; \max_{P\in\mathcal{P}} \; \bar{e}_p(d).$$

It is obvious from the definitions of d_* and d^* that

$$\max_{P\in\mathcal{P}} \; \bar{e}_P(d^*) \le \max_{P\in\mathcal{P}} \; \bar{e}_P(d_*)$$

and, at the same time,

$$\max_{P\in\mathcal{P}} \; e_P(d_*(x);x) \le \max_{P\in\mathcal{P}} \; e_P(d^*(x);x) \qquad \text{for all} \;\; x\in\mathbf{X}.$$

These properties seem to be contradictory, but they are not as

$$\max_{P\in\mathcal{P}} \; \bar{e}_P(d) = \max_{P\in\mathcal{P}} \; \sum_{x} \; e_P(d(x);x) \le \; \sum_{x} \max_{P\in\mathcal{P}} \; e_P(d(x);x).$$

Example 1 in the Appendix shows that these inequalitites are strict, generally. It can be seen that d_* in this example is the optimal decision function with respect to P_2 and d_3 is the optimal one with respect to P_1. So $d^* = d_2$ is optimal with respect to none of the distributions in $\mathcal{P}$. It is not difficult to find other examples where d_* again is not optimal with respect to any distribution in $\mathcal{P}$.

The question we want to answer is the following: What are the conditions under which d^* (resp. d_*) is an optimal decision function for some $P \in \mathcal{P}$? If such P exists, it can be used as a representation of the minimax decision function. This distribution can be, in fact, more appropriate for practical use than the decision function (see e.g. [2]). Let us denote $\mathcal{P}^* \subset \mathcal{P}$ the set of all distributions from $\mathcal{P}$ with respect to which the minimax decision function d^* is optimal. So, the first question is whether $\mathcal{P}^*$ is empty or not.

This problem will be solved for d^* (although similar conclusions can be done for d_* as well) and for $\mathcal{P}$ being a convex linear polyedr.

2. EXPLICIT SOLUTION OF THE PROBLEM

It can be seen from the form of the expected error of a decision function that if a distribution P_α is a linear combination of distributions P_1 and P_2 then the error $e_{P_\alpha}(d(x);x)$ of every decision $d(x)$ and the expected error $\bar{e}_p(d)$ of every decision function d is a linear combination of the decision errors $e_{P_1}(d(x);x)$ and $e_{P_2}(d(x);x)$ and the expected errors of the decision functions $\bar{e}_{P_1}(d)$ and $\bar{e}_{P_2}(d)$ respectively. In general, taking into account the convex polyedr with k vertices in $(k-1)$-dimensional space, we have

$$e_{P_\alpha}(d(x);x) = \sum_{i=1}^{k} \alpha_i e_{P_i}(d(x);x)$$

and

$$\bar{e}_{P_\alpha}(d) = \sum_{i=1}^{k} \alpha_i \bar{e}_{P_i}(d), \tag{1}$$

for

$$P_\alpha = \sum_{i=1}^{k} \alpha_i P_i, \quad \alpha = (\alpha_1, \ldots, \alpha_k), \quad \alpha_i \geq 0, \quad \sum_{i=1}^{k} \alpha_i = 1.$$

What does the optimality of a decision function mean? The decision function d_P is optimal with respect to the distribution P if

$$\bar{e}_P(d_P) \leq \bar{e}_P(d)$$

for all decision functions d. Therefore the question whether the minimax decision function d^* (as mentioned above we could consider d_* and $e_P(d(x);x)$ instead of d^* and $\bar{e}_P(d)$ as well) is optimal for some P_α is equivalent to the problem whether there exists α satisfying

$$\bar{e}_{P_\alpha}(d^*) \leq \bar{e}_{P_\alpha}(d)$$

for every decision function d . Rewriting this inequality using (1), we get

$$\sum_{i=1}^{k} \alpha_i \bar{e}_{P_i}(d^*) \leq \sum_{i=1}^{k} \alpha_i \bar{e}_{P_i}(d).$$

Since it is a convex linear combination, the $\alpha^{'}s$ have to satisfy in addition the condition

$$\sum_{i=1}^{k} \alpha_i = 1, \quad \alpha_i \geq 0, \quad i = 1, ..., k.$$

In the case of $k = 2$ this problem can be solved explicitly. Let us set $\alpha_1 = \beta$, then $\alpha_2 = 1 - \beta$ and we have the inequality

$$\beta \bar{e}_{P_1}(d^*) + (1-\beta)\bar{e}_{P_2}(d^*) \leq \beta \bar{e}_{P_1}(d) + (1-\beta)\bar{e}_{P_2}(d),$$

or equivalently

$$\beta[\bar{e}_{P_1}(d^*) - \bar{e}_{P_1}(d)] \leq (1-\beta)[\bar{e}_{P_2}(d) - \bar{e}_{P_2}(d^*)],$$

and therefore

$$\beta[\bar{e}_{P_1}(d^*) - \bar{e}_{P_1}(d) + \bar{e}_{P_2}(d) - \bar{e}_{P_2}(d^*)] \leq \bar{e}_{P_2}(d) - \bar{e}_{P_2}(d^*). \tag{2}$$

Theorem 1 is an immediate consequence of this inequality.

Theorem 1. Let $\mathcal{P} = \{P_\beta = \beta P_1 + (1-\beta)P_2,\ \beta \in [0,1]\}$ and d^* be the global minimax decision function. Let us denote

$$\beta(d) = \frac{\bar{e}_{P_2}(d) - \bar{e}_{P_2}(d^*)}{\bar{e}_{P_1}(d^*) - \bar{e}_{P_1}(d) + \bar{e}_{P_2}(d) - \bar{e}_{P_2}(d^*)}.$$

If there exists β^* satisfying inequalities

$$\beta^* \leq \beta(d)$$

for all d satisfying $\bar{e}_{P_1}(d) - \bar{e}_{P_2}(d) < \bar{e}_{P_1}(d^*) - \bar{e}_{P_2}(d^*)$, and

$$\beta^* \geq \beta(d)$$

for all d satisfying the opposite inequality, then $P_{\beta^*} \in \mathcal{P}^* \subset \mathcal{P}$.

Remark. It should be stressed that the respective β has to meet as many inequalities as there are different decision functions with the exception of those for which the denominator is equal to zero. In this case

$$\bar{e}_{P_1}(d^*) - \bar{e}_{P_1}(d) = \bar{e}_{P_2}(d^*) - \bar{e}_{P_2}(d)$$

and therefore

$$\bar{e}_{P_1}(d) \geq \bar{e}_{P_1}(d^*) \quad \text{and} \quad \bar{e}_{P_2}(d) \geq \bar{e}_{P_2}(d^*)$$

for d^* being the global minimax decision function. Therefore the inequality (2) holds for all β.

3. SOLUTION VIA THE LINEAR PROGRAMMING METHODS

In general, however, such a simple optimality criterion (as set forth in Theorem 1) cannot be found. Let us denote $\mathcal{D}$ the set of all decision functions from $\mathbf{X}$ to $\mathbf{Y}$ excluding d^* (let us notice that $\mathcal{D}$ is finite since both $\mathbf{X}$ and $\mathbf{Y}$ are finite) and $J = card(\mathcal{D})$. We can index decision functions in $\mathcal{D}$ by numbers $1, ..., J$ and we get the following problem: To find out whether the system of J inequalities

$$\sum_{i=1}^{k} \alpha_i [\bar{e}_{P_i}(d_j) - \bar{e}_{P_i}(d^*)] \geq 0, \qquad j = 1, \ldots, J$$

and the equality

$$\sum_{i=1}^{k} \alpha_i = 1,$$

has at least one solution such that $\alpha_i \geq 0, \quad i = 1, \ldots, k$.

This problem can be reformulated in terms of linear programming: To minimize an arbitrary constant function under the conditions stated above. It can be easily solved using the simplex algorithm (see e.g. [1]). Using slack variables $\alpha_{k+1}, \ldots, \alpha_{k+J}$, we get equality constraints

$$\sum_{i=1}^{k} \alpha_i [\bar{e}_{p_i}(d_j) - \bar{e}_{p_i}(d^*)] - \alpha_{k+j} = 0, \qquad j = 1, \ldots, J,$$

and

$$\sum_{i=1}^{k} \alpha_i = 1,$$
$$\alpha_i \geq 0, \qquad i = 1, \ldots, k + J.$$

In order to determine some basic solution we use an artificial variable α_{k+J+1}. The problem has then the form

$$\min \quad w\,\alpha_{k+J+1} \tag{3}$$

(where w is some large constant) under conditions

$$\begin{aligned} \sum_{i=1}^{k} \alpha_i[\bar{e}_{P_i}(d^*) - \bar{e}_{P_i}(d_j)] + \alpha_{k+j} = 0, \qquad & j = 1, \ldots, J, \\ \sum_{i=1}^{k} \alpha_i + \alpha_{k+J+1} = 1, & \\ \alpha_i \geq 0, \qquad i = 1, \ldots, k+J+1. & \end{aligned}$$

From this form we can easily get the basic solution

$$(\alpha_{k+1}, \ldots, \alpha_{k+J}, \alpha_{k+J+1}) = (0, \ldots, 0, 1).$$

This solution, however, is not feasible. Applying the simplex method, we get eventually the optimal solution of this problem.

It is well known (see e.g. [1]) that if this solution involves the artificial variable, a feasible solution of the primal problem does not exist. In the other case, the optimal solution of the problem (3) is a feasible solution of the primal problem and we can find all feasible solutions $\alpha(l)$, $l = 1, ..., L$ (for some $L \leq \frac{(k+J)!}{(J+1)!(k-1)!}$), of the primal problem (see [1] again).

This general result obtained by linear programming methods can be interpreted for our purpose in this way: If the solution involves the artificial variable α_{k+J+1}, the set $\mathcal{P}^*$ is empty. Otherwise the set $\mathcal{P}^*$ is a convex polyedr with vertices $P_{\alpha(l)}$, $l = 1, ..., L$, where $\alpha_1, ..., \alpha_k$ are certain coordinates of $\alpha(l)$.

But what shall we do if $\mathcal{P}^*$ is the empty set? One possible answer will be offered in the next section.

4. APPROXIMATION OF DISTRIBUTIONS IN $\mathcal{P}$

In order to solve this problem we have at first to define an *f-divergence*, which is used to measure the dissimilarity of two distributions. The f-divergence of probabilities P and Q is defined in the discrete case for every function $f(u)$ convex on $(0, \infty)$ and strictly convex at the point $u = 1$ as

$$D_f(P, Q) = \sum_{x \in \mathbf{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

(setting

$$Q(x) f\left(\frac{P(x)}{Q(x)}\right) = P(x) \lim_{u \to \infty} \frac{f(u)}{u}$$

for such x that $Q(x) = 0$ and $P(x) > 0$).

Let us return to the problem stated above. For every decision function d we can find a set of distributions $\mathcal{P}_d$ this decision function is optimal for. At this moment, we consider all joint probability distributions defined for the pair of variables XY. Therefore it can be easily seen that $\mathcal{P}_d \neq \emptyset$ for any d (it is enough to consider the distribution $P_d(x, y) = \frac{1}{card\mathbf{X}}$ if $d(x) = y$ and $P_d(x, y) = 0$ otherwise). Consider d^* and the set $\mathcal{P}_{d^*}$ of distributions which have the same optimal decision function d^*. Any of these distributions can be used to represent the decision function d^*. But some of them are "too distant" from the original set $\mathcal{P}$ of the distributions having the minimax decision function d^*. For the purpose of the representation of d^* it

seems to be reasonable to use such distribution $P^* \in \mathcal{P}_{d^*}$, which is the "closest" possible to the distributions from $\mathcal{P}$, i.e. an approximation of distributions from the set $\mathcal{P}$ in the set $\mathcal{P}_{d^*}$. It can be done using the barycenter concept introduced by Perez in [3].

Let us consider sets $\mathcal{P}$ and $\mathcal{Q}$ (not necessarilly different). Distribution Q^* will be called a *D_f-barycenter of a set $\mathcal{P}$ with respect to a set $\mathcal{Q}$*, if

$$Q^* \in \arg\min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}} D_f(P,Q).$$

Distribution Q^* then has the following characteristics:

1. The minimax decision function is optimal with respect to it, i.e.

$$e_{Q^*}(d^*) \le e_{Q^*}(d)$$

for all possible decision functions d.

2. The divergence $D_f(P,Q^*)$ is minimal for the least favourable $P \in \mathcal{P}$, i.e.

$$\max_{P\in\mathcal{P}} D_f(P,Q^*) \le \max_{P\in\mathcal{P}} D_f(P,Q)$$

for every $Q \in \mathcal{Q}$.

The practical construction of the D_f-barycenter is another problem whose solution is quite difficult and exceeds the framework of this paper. But the whole procedure (decision whether there exists, or not, any $P \in \mathcal{P}$ with respect to which d^* is optimal, the construction – in the latter case – of the set $\mathcal{P}_{d^*}$ and finding the barycenter of $\mathcal{P}$ in $\mathcal{P}_{d^*}$) is demonstrated in Example 2 in the Appendix.

Barycenter of the set $\mathcal{P}$ with respect to the set $\mathcal{Q}$ need not be unique, which is shown in Example 3 of the Appendix.

APPENDIX

Example 1. Let the class $\mathcal{P}$ consist of the two following distributions

$$\begin{aligned} P_1(0,0) &= \tfrac{1}{6}, & P_2(0,0) &= \tfrac{1}{3},\\ P_1(0,1) &= \tfrac{1}{4}, & P_2(0,1) &= \tfrac{1}{8},\\ P_1(1,0) &= \tfrac{1}{4}, & P_2(1,0) &= \tfrac{5}{12},\\ P_1(1,1) &= \tfrac{1}{3}, & P_2(1,1) &= \tfrac{1}{8}. \end{aligned}$$

Knowing the value of the variable X, we want to decide about the value of the variable Y. Using the local minimax rule we get

$$d_*(0) = 0 \ \text{ and } \ \mathrm{d}_*(1) = 0.$$

Defining the other functions

$$\begin{aligned} d_1(0) &= 0, & d_1(1) &= 1,\\ d_2(0) &= 1, & d_2(1) &= 0,\\ d_3(0) &= 1, & d_3(1) &= 1, \end{aligned}$$

we can compute corresponding errors

$$\begin{aligned} e_{P_1}(d_1) &= \tfrac{1}{2}, & e_{P_2}(d_1) &= \tfrac{13}{24},\\ e_{P_1}(d_2) &= \tfrac{1}{2}, & e_{P_2}(d_2) &= \tfrac{11}{24},\\ e_{P_1}(d_3) &= \tfrac{5}{12}, & e_{P_2}(d_3) &= \tfrac{3}{4},\\ e_{P_1}(d_*) &= \tfrac{7}{12}, & e_{P_2}(d_*) &= \tfrac{1}{4}. \end{aligned}$$

It is obvious that not d_* but d_2 is the global minimax decision function.

In Section 4 we have defined the f-divergence. There is a large class of functions satisfying the requirements stated there, i.e. we can define a lot of divergences (see e.g. [4]). In our examples we will use only one of them – the *total variation* ($f(u) = |u-1|$)

$$V(P,Q) = \sum_{x\in\mathbf{X}} |P(x) - Q(x)|.$$

Example 2. Let us consider following convex set $\mathcal{P}$ of distributions

$$\begin{aligned} P_\alpha(0,0) &= \tfrac{1}{12} + \tfrac{5}{12}\alpha, \\ P_\alpha(0,1) &= \tfrac{5}{24} - \tfrac{1}{8}\alpha, \\ P_\alpha(1,0) &= \tfrac{1}{3}, \\ P_\alpha(1,1) &= \tfrac{3}{8} - \tfrac{7}{24}\alpha, \quad \alpha \in [0,1]. \end{aligned}$$

It is not difficult to find out that $d_2 \equiv 1$ is the optimal decision function for $P_\alpha, \alpha \in [0, \frac{1}{7}], d_3(0) = 1, d_3(1) = 0$ is the optimal one for $P_\alpha, \alpha \in [\frac{1}{7}, \frac{3}{13}]$ and $d_1 \equiv 0$ is the optimal one for $P_\alpha, \alpha \in [\frac{3}{13}, 1]$. But the minimax decision function is

$$d^*(0) = 0, \quad d^*(1) = 1$$

and so it is not optimal for any $\alpha \in [0,1]$.

Let us consider a set

$$\begin{aligned} \mathcal{P}^* = \{P: \quad & P(0,0) = p_1, P(0,1) = p_2, P(1,0) = p_3, P(1,1) = 1 - p_1 - p_2 - p_3, \\ & p_1 \geq p_2, p_1 + p_2 + p_3 \leq 1\}. \end{aligned} \tag{4}$$

The decision function d^* is the optimal one with respect to the distributions from this set.

The total variation of the distributions $P_\alpha \in \mathcal{P}$ and $P \in \mathcal{P}^*$ is

$$V(P_\alpha, P) = |\frac{1}{12} + \frac{5}{12}\alpha - p_1| + |\frac{5}{24} - \frac{1}{8}\alpha - p_2| + |\frac{1}{3} - p_3| + |-\frac{5}{8} - \frac{7}{24}\alpha + p_1 + p_2 + p_3|.$$

It is obvious that to determine

$$\max_{P_\alpha\in\mathcal{P}} V(P_\alpha, P) = \max_{\alpha\in[0,1]} V(P_\alpha, P)$$

only values in boundary points of the interval [0,1] are important (for $V(P_\alpha, P)$ is a convex function of the variable α). So, we will be interested only in maxima of the values $V(P_0, P)$ and $V(P_1, P)$. It is too laborious and not very interesting to determine the subsets of $[0,1]^3$, where one or the other value is maximal. Therefore, it is not done here, but only the result is stated. We have to determine values of parameters p_1, p_2 and p_3, which the $\max\{V(P_0, P),\ V(P_1, P)\}$ is minimal for. These values are

$$p_1 = p_3 = \frac{11}{36}, \quad p_2 = \frac{1}{12}$$

and the maximum value is

$$\max\{V(P_0, P),\ \ V(P_1, P)\} = \frac{4}{9}.$$

So the barycenter of the set $\mathcal{P}$ with respect to the set $\mathcal{P}^*$ is the distribution P^*:

$$\begin{aligned} P^*(0,0) &= \tfrac{11}{36}, \\ P^*(0,1) &= \tfrac{1}{12}, \\ P^*(1,0) &= \tfrac{11}{36}, \\ P^*(1,1) &= \tfrac{11}{36} \end{aligned}$$

and the following equality holds

$$\max_{P\in\mathcal{P}}\ V(P_\alpha, P^*) = \frac{4}{9}.$$

Example 3. Let us consider a convex set $\mathcal{P}$ of distributions

$$\begin{aligned} P_\alpha(0,0) &= \tfrac{1}{6} + \tfrac{23}{48}\alpha, \\ P_\alpha(0,1) &= \tfrac{5}{12} - \tfrac{19}{48}\alpha, \\ P_\alpha(1,0) &= \tfrac{1}{6} + \tfrac{7}{48}\alpha, \\ P_\alpha(1,1) &= \tfrac{1}{4} - \tfrac{11}{48}\alpha, \qquad \alpha \in [0,1]. \end{aligned}$$

It is clear, that $d_2 \equiv 1$ is optimal for $P_\alpha, \alpha \in [0, \frac{4}{15}]$ (by the way $d_* \equiv d_2$) and $d_1 \equiv 0$ is optimal for $P_\alpha, \alpha \in [\frac{4}{15}, 1]$, while

$$d^*(0) = 0, \quad d^*(1) = 1.$$

A set of distributions with respect to which d^* is optimal has the form (4) again. Using the same procedure as in Example 2, we will find out, that the barycenter of the set $\mathcal{P}$ with respect to $\mathcal{P}^*$ is an arbitrary distribution from the set

$$\begin{aligned} \mathcal{P}_0^* = \{P:\ \ & P(0,0) = p_1, P(0,1) = p_2, P(1,0) = \tfrac{31}{48} - p_1, P(1,1) = \tfrac{17}{48} - p_2, \\ & p_1 \geq p_2 + \tfrac{7}{24}\}. \end{aligned}$$

(Received September 25, 1991.)

REFERENCES

[1] S. I. Gass: Linear Programming – Methods and Applications. McGraw-Hill, New York 1969.
[2] R. Jiroušek: A survey of methods used in probabilistic expert systems for knowledge integration. Knowledge Based Systems *3* (1990), 7–12.
[3] A. Perez: "Barycenter" of a set of probability measures and its application in statistical decision. In: Compstat 1984 (Proceedings in Computational Statistics), Wien, Physica-Verlag 1984, pp. 154–159.
[4] I. Vajda: Theory of Statistical Inference and Information. Kluwer, Dordrecht – Boston 1989.

RNDr. Jiřina Vejnarová CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation – Czechoslovak Academy of Sciences), Pod vodárenskou veží 4, 182 08 Praha 8. Czechoslovakia.