77-10

Capacity and Efficiency of Decision Functions

JAMES R. BITNER

Abstract—Two criteria for comparing the performance of decision functions are discussed. The first is the *capacity* as defined by Cover. We extend his definition to allow the calculation of the capacity of hyperboxes (interval complexes), which are shown to have asymptotically lower capacity than hyperplanes. The second criterion is the *efficiency*. We calculate the efficiency for polynomial functions. The efficiency sharply decreases as the degree of the polynomial increases. This calculation is also done for hyperboxes, which are shown to be more efficient than hyperplanes.

Index Terms—Capacity, decision functions, dichotomy, efficiency, hyperboxes, hyperplanes, interval complexes, Φ-machine.

I. INTRODUCTION

The *capacity* of the class of rth-degree polynomial decision functions has been defined by Cover ([1], [2], [11], see also [6]) as follows: Given N points in a d-dimensional space, calculate the fraction of the 2^N dichotomies that are achievable using polynomials of a given degree (this fraction can also be viewed as the *dichotomization probability*; i.e., the probability that a random dichotomy is achievable). As N increases and d is held fixed, this fraction is first close to one, then drops quickly to zero. The point where this drop occurs is defined as the capacity; it measures the number of points a function can reasonably handle in a given number of dimensions. For rth-degree polynomials, the capacity was found to be $2\binom{d+r}{d}$. Note that this is twice the number of adjustable coefficients.

Cover's work raises several questions: Is capacity a meaningful quantity for other classes of decision functions? If so, what is their capacity and how does it compare with that of the class of polynomial functions? We discuss these questions for the class of

Manuscript received January 27, 1976; revised December 14, 1976. This work was supported in part by the National Science Foundation under Grants GJ-41538 and DCR 74-03514.

The author was with the University of Illinois at Urbana-Champaign, Urbana, IL. He is now with the Department of Computer Science, University of Texas, Austin, TX 78712.

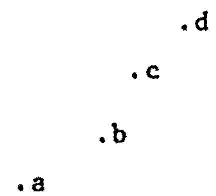

Fig. 1.

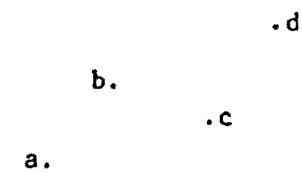

Fig. 2.

hyperboxes (interval complexes), an integral part of the variable-valued logic system of Michalski [4], [5]. We also consider the efficiency, as defined by Zagoruiko [3], of this class of functions and of polynomial decision functions.

II. Capacity of Hyperboxes

A hyperbox (or interval complex [4], [5]) is the region in d-space defined by $\{\bar{x} = (x_1, \cdots, x_d) | a_i \le x_i \le b_i\}$. Notice that the sides of the hyperbox must be *parallel* to the axes. Since a hyperbox has $2d$ coefficients, and since the capacity of any polynomial decision function equals twice the number of adjustable coefficients, it might be tempting to *define* the capacity of a hyperbox to be $4d$. However, such a definition is clearly unacceptable because capacity would no longer have its intuitive meaning of "the greatest number points a function can handle in a given number of dimensions." The capacity must be redetermined by calculating the probability that a random dichotomy of N points in d-space is achievable by a hyperbox.

To begin the analysis, we first note that the number of dichotomies achievable with hyperboxes varies with the location of the points. Consider Figs. 1 and 2. In Fig. 1 the dichotomies with a and c in one class and b and d in the other cannot be achieved, but in Fig. 2 all dichotomies are possible. It would be natural to restrict the points from lying in some improbable degenerate configuration as Cover did for polynomial functions by requiring that the points be in general position. However, this is not possible; as long as d is "northeast" of c, the dichotomies in Fig. 1 will still be unachievable.

To cope with this problem, we need to make only two very general assumptions. First, that the points are independently chosen from an arbitrary continuous probability density, and second, that the value chosen for one coordinate of a point is independent of its other coordinates.

Now, to determine if a given dichotomy is achievable, index through the dimensions. A class-2 point (called a "two") is said to be *active* if, in each dimension we have examined, the two has occurred between the leftmost one and the rightmost one in that dimension. After all dimensions have been examined, there exists a hyperbox containing all the ones and no twos if all twos are now inactive. Since the dichotomy is achievable if either the ones or the twos can be contained in a hyperbox, we must repeat this process, with the roles of ones and twos interchanged.

From this procedure it can be seen that only the *order* of the points (from left to right) in each dimension determines if a dichotomy can be achieved; the number of dichotomies is independent of the exact coordinates of the points, as long as their relative order remains the same. Therefore, a distribution of N points can be described by d sequences of length N, with the ith sequence listing the points from left to right in dimension i. Since we assume the N points are independently chosen, each of the $N!$ orderings (sequences) is *equally likely*. This holds for *any* probability density. Further, sequences for different dimensions are *independent*, and since we desire the probability of achieving a random dichotomy, each point is equally likely to belong to class 1 or class 2.

We now describe the calculation of Prob (N points in a d-dimensional space can be dichotomized), the dichotomization probability. Since the probability that k of the N points are ones equals $\binom{N}{k}/2^N$, the dichotomization probability is

$$(1/2^N) \sum_{k=0}^{N} \binom{N}{k} \times \text{Prob}\,(k \text{ ones and } N-k \text{ twos can be dichotomized}). \quad (1)$$

Define an *outer* one to be a one that is *not* between the leftmost and rightmost twos in a given dimension. All outer ones and twos will become inactive (some may be already). For $m = 1, \cdots, d$, we do the following calculation for all $i' \le k$ and $j' \le N-k$.

$$\begin{aligned}&\text{Prob}\,(i' \text{ inactive ones and } j' \text{ inactive twos after } m \text{ dimensions})\\ &= \sum_{i \le i'} \sum_{j \le j'} \sum_{x,y} \text{Prob}\,(x \text{ outer ones, } y \text{ outer twos})\\ &\quad \times \text{Prob}\,(i'-i \text{ out of } x \text{ outer ones are active})\\ &\quad \times \text{Prob}\,(j'-j \text{ out of } y \text{ outer twos are active})\\ &\quad \times \text{Prob}\,(i \text{ inactive ones and } j \text{ inactive twos after}\\ &\qquad m-1 \text{ dimensions}) \quad (2)\end{aligned}$$

where, if $m = 1$, the last probability equals one if $i = j = 0$ and zero otherwise. The probability in the summand of (1), which equals the probability that all the ones or all the twos are inactive after d dimensions, can easily be determined from the results of (2) with $m = d$.

The calculation of (2) has two parts. We first describe the calculation of $P(x$ outer ones, y outer twos). There are six cases.

Case 1 ($x = 0$): To get exactly y outer twos, the sequence must be $2^a 1 z 1 2^b$,[1] where $a + b = y$, $a > 0$, $b > 0$, and z is any string of $k-2$ ones and $N-k-a-b$ twos. There are $(y-1)\binom{N-y-2}{k-2}$ ways of choosing the sequence. Hence the probability is $(y-1)\binom{N-y-2}{k-2}/\binom{N}{k}$, since there are $\binom{N}{k}$ strings with k ones.

Case 2 ($y = 0$): This is the same as Case 1 with the roles of the ones and twos reversed. We have probability equal to

$$\frac{(x-1)\binom{N-x-2}{k-x}}{\binom{N}{k}}.$$

Case 3 ($0 < x < k$ and $0 < y < N-k$): The sequence must be $1^x 2 z 1 2^y$ or $2^y 1 z 2 1^x$, where z has $k-x-1$ ones out of $N-x-y-2$ positions. Therefore, the probability is

$$\frac{2\binom{N-x-y-2}{k-x-1}}{\binom{N}{k}}.$$

Case 4 ($x = k$ and $0 < y < N-k$): This case is impossible; when all the ones are outer, either all twos must be outer or they all must be inner.

Case 5 ($0 < x < k$ and $y = N-k$): Similar to Case 4 and also impossible.

[1] The exponent in regard to repetition 2^a means we have a consecutive twos.

TABLE I
Table of Probabilities that N Points Can Be Dichotomized in d Dimensions

	d (Number of dimensions)								
N (number of points)	1	2	3	4	5	6	7	8	9
4	0.8750	0.9896	0.9991	0.9999	0.9999	0.9999			
5	0.6875	0.9396	0.9913	0.9988	0.9998	0.9999			
6	0.5000	0.8358	0.9615	0.9920	0.9984	0.9997			
7	0.3437	0.6955	0.8963	0.9697	0.9917	0.9978			
8	0.2266	0.5459	0.7955	0.9222	0.9730	0.9911			
9	0.1445	0.4083	0.6718	0.8456	0.9348	0.9742			
10	0.0898	0.2935	0.5420	0.7448	0.8730	0.9415			
11	0.0547	0.2042	0.4201	0.6804	0.7891	0.8891			
12	0.0327	0.1383	0.3147	0.5143	0.6894	0.8172			
13	0.0193	0.0915	0.2290	0.4059	0.5828	0.7297			
14	0.0112	0.0594	0.1625	0.8112	0.4779	0.6329			
15	0.0065	0.0379	0.1129	0.2327	0.3811	0.5338			
16	0.0037	0.0238	0.0770	0.1702	0.2965	0.4388			
17	0.0021	0.0148	0.0516	0.1220	0.2257	0.3522	0.4858	0.6113	0.7186
18	0.0012	0.0091	0.0341	0.0860	0.1685	0.2768	0.3998	0.5241	0.6381
19	0.0007	0.0055	0.0223	0.0598	0.1236	0.2134	0.3224	0.4401	0.5552

Case 6 ($x = k$ and $y = N - k$): Here only two configurations ($1^x 2^y$ and $2^y 1^x$) are possible, so the probability is $2/\binom{N}{k}$.

Finally, what is Prob ($i' - i$ out of x outer ones are active)? We must choose $i' - i$ of the x outer ones and $k - i'$ of the $k - x$ inner ones to be active (there are exactly $k - i$ *active* ones in all). Since there are $\binom{x}{i'-i} \cdot \binom{k-x}{k-i'} = \binom{x}{i'-i}\binom{k-x}{i'-x}$ such choices and a total of $\binom{k}{i}$ ways of arbitrary placement, the probability is

$$\frac{\binom{x}{i'-i} \cdot \binom{k-x}{i'-x}}{\binom{k}{i}}.$$

Prob ($j' - j$ out of y outer twos are active) is calculated in a similar way and equals

$$\frac{\binom{y}{j'-j}\binom{N-k-y}{j'-y}}{\binom{k}{i}}.$$

This completes the description. A program was written to exactly calculate the dichotomization probabilities. The results are shown in Table I.

We are interested in the points where these curves equal 1/2. These points can be closely approximated using linear interpolation (the curves are fairly linear in this region). The results are shown in Table II.

The slope is the difference between successive values of $C(d)$. We make three observations from this table.

1) The dichotomization probability has the same form as that for polynomial functions; it drops off quickly after the capacity has been reached. Therefore, capacity *is* a meaningful quantity for hyperboxes.

2) The capacity of a hyperbox will be asymptotically less than that of the hyperplane, even though the hyperbox has nearly twice as many adjustable coefficients. However, for small values (this may be of practical interest) the hyperbox has higher capacity.

3) The slope of $C(d)$ decreases as d increases. Somehow, later dimensions are not as "useful" as earlier ones. Note that every dimension adds a constant to the capacity of the hyperplane; all dimensions have the same "usefulness."

The slope appears to approach a limit around 1.4, indicating the capacity approaches a linear asymptote having slope 1.4. The point $C(8) = 18.2869$ determines the y-intercept, giving a capacity of $1.4d + 7$ for large d. To obtain more evidence, a second program was written that approximates the probabilities by randomly generating dichotomies and testing whether they are achievable. For $N = 24$ and $d = 12$, the estimated probability was 0.480, agreeing nicely with the predicted $1.4(12) + 7 = 23.8$.

TABLE II
Value for $C(d)$, the Capacity of d-Dimensional Hyperboxes

d	C(d)	slope	Capacity of a hyperplane
1	6.00000		4
		2.33358	
2	8.33358		6
		2.01096	
3	10.34454		8
		1.78738	
4	12.13192		10
		1.65740	
5	13.78932		12
		1.56665	
6	15.35597		14
		1.48884	
7	16.84481		16
		1.44209	
8	18.28690		18

It is not clear why hyperboxes should have such a low capacity compared to their $2d$ adjustable coefficients. An argument against the results is that a hyper*cube* is an equimetric surface under the metric $d(\bar{x},\bar{y}) = \max_i |x_i - y_i|$, so we might expect it to have the same capacity as a hypersphere (an equimetric under the Euclidean metric). A hyperbox would then have a much higher capacity than a hypercube, since we can choose d "radii" instead of one.

However, there are intuitive arguments supporting the results: Hyperspheres always can achieve a linear dichotomy (a large class) by making the radius very large. Hyperboxes cannot since the sides must be parallel to the axes. Another argument is that the dimensions are considered independently in determining if a point is in a hyperbox, while for hyperplanes and hyperspheres, the dimensions are interrelated. In general, we would expect a method that makes use of this interrelation to be more powerful.

III. Efficiency of Polynomial Decision Functions

We now discuss the idea of *efficiency*, defined by Zagoruiko [3] to be the ratio of the capacity to the product of the amount of storage and the computational effort required by the function. Zagoruiko incorrectly calculated the efficiency of polynomial decision functions. We now correctly calculate this efficiency and discuss its significance.

We determine the number of operations required to evaluate an rth-degree polynomial in d variables, using an extension of Horner's rule [9] which evaluates the polynomial in the minimum time possible. The rule is recursively defined as follows.

TABLE III
The Efficiency of Various Decision Functions in d Dimensions

d	hyperbox	r=1	r^{th} degree polynomial r=2	r=3
1	3.42857	.68493	.25685	.18265
2	1.36059	.34247	.10274	.06342
3	.78816	.22831	.05854	.03135
4	.53327	.17123	.03862	.01808
5	.39398	.13699	.02766	.01143
10	.15484	.06227	.01245	.00415
15	.09276	.04280	.00570	.00121
20	.06558	.03261	.00326	.00051

For $d = 1$, use the standard form of Horner's rule. For $d > 1$, collect all terms with the same power of x_1 to obtain $f(x_1, \cdots, x_d) = a_r x_1^r + x_1^{r-1} P_1(x_2, \cdots, x_d) + \cdots + x_1 P_{r-1}(x_2, \cdots, x_d) + P_r(x_2, \cdots, x_d)$, where $P_i(x_2, \cdots, x_d)$ is an ith-degree polynomial in $d - 1$ variables. Each P_i is then evaluated recursively. To evaluate f, the values obtained for the P_i are viewed as coefficients of an rth-degree polynomial in x_1, which is evaluated using Horner's rule. This method requires $\binom{d+r}{r} - 1$ additions and $\binom{d+r}{r} - 1$ multiplications to evaluate an rth-degree polynomial in d variables.

The storage needed is $\binom{d+r}{r}$ locations for the coefficients. In addition, d locations are required to store intermediate results if $r > 1$. This gives an efficiency (E) of:

$$E = \frac{2\binom{d+r}{r}}{\left[\binom{d+r}{r} + d\right]\left[\binom{d+r}{r} - 1\right](t_A + t_M)}, \qquad \text{for } r > 1$$

and

$$E = \frac{2\binom{d+1}{1}}{\binom{d+1}{1}\left[\binom{d+1}{1} - 1\right](t_A + t_M)} = \frac{2}{d(t_A + t_M)}, \qquad \text{for } r = 1$$

where t_A and t_M are the times required to add and multiply. As r and d increase, $\binom{r+d}{r}$ increases very rapidly and quickly dominates the expression. Thus, for large r and d, we have

$$E \approx \frac{2\binom{d+r}{r}}{\binom{d+r}{r}\binom{d+r}{r}(t_A + t_M)} = \frac{2}{\binom{d+r}{r}(t_A + t_M)}.$$

So as r and d increase, the functions rapidly become less and less efficient. Thus, even though high-degree polynomials have a large capacity, it is outweighed by their even larger complexity, making the hyperplane the most desirable function of this class.

IV. Efficiency of Hyperboxes

To calculate the efficiency of hyperboxes, we first determine the storage and time requirements. Suppose the feature space is discrete (as it always is using interval complexes) and that the following representation is used: A feature (x_i) which may take on v different values and has value j is represented by setting the jth bit in a string of v bits. The constraint $a_i \le x_i \le b_i$ is represented by setting bits a_i through b_i in a string of v bits. Testing whether a point satisfies a given constraint is accomplished by one "and" operation.

We require d locations to store the constraints. To verify a point is in the hyperbox requires d "and" operations, since all d constraints must be tested. If a point is not in the hyperbox, we stop testing as soon as *any* constraint is violated. Assuming that half the points to be classified lie in the hyperbox and that, on the average, half of the constraints $\lfloor (d+1)/2 \rfloor$ are examined in classifying points *not* in the hyperbox, $\frac{1}{2}d + \frac{1}{2}\lfloor (d+1)/2 \rfloor = (3d+1)/4$ "and's" will be required. This gives the efficiency as

$$\frac{1.4d + 7}{d\left(\frac{3d+1}{4}\right) t_{\text{and}}} = \frac{5.6d + 28}{d(3d+1)t_{\text{and}}}.$$

Table III compares this with polynomial discriminant functions, where t_A, t_M, and t_{and} are the execution times of the appropriate instruction on an IBM 360/75. The quantity t_{and} also contains the time for a conditional branch. The efficiency for the hyperbox with $d \le 5$ is calculated using the exact values. For $d > 5$, the approximation $C(d) \approx 1.4d + 7$ was used.

Actually, the hyperbox's efficiency is even higher in practice. After synthesizing an interval complex formula, we usually find that only a fraction α of the dimensions have constraints imposed on them. In this case, both the storage and time requirements are multiplied by α, and the total efficiency is divided by α^2. Typically, $\alpha = \frac{1}{3}$, and the efficiency is 9 times greater than that shown in the table.

V. Conclusion

Our main task has been to calculate the *capacity* (as defined by Cover) of the class of *hyperboxes*, an important component of the variable-valued logic system of Michalski. The capacity was found to be $1.4d + 7$ for large d. This calculation also revealed three interesting facts. First, the probability that a dichotomy of N points can be achieved by a hyperbox has the same behavior as that for hyperplanes. That is, it is close to one for a small number of points, then drops quickly and is close to zero. This, of course, was *not known a priori* (the probability might slope very slowly down toward zero). Therefore, capacity *is* a meaningful quantity for hyperboxes (a figure very different from polynomial functions), indicating the robustness of Cover's definition.

Second, the capacity of $1.4d + 7$ is surprisingly small. Since the capacity of polynomial functions equals twice the number of adjustable coefficients, we would expect a capacity of $4d$. The hyperbox somehow makes "less use" of its degrees of freedom than a polynomial function.

Third, the slope of capacity function decreases (approaching 1.4 as a limit). Therefore, increasing the number of dimensions gives diminishing returns, where earlier dimensions are more useful than later ones.

We also discussed the *efficiency* as defined by Zagoruiko. We corrected his calculation for the efficiency of polynomial functions and found that the efficiency decreased rapidly as the degree of the polynomial increased. We extended this calculation to hyperboxes and found them to be *more efficient* than hyperplanes.

ACKNOWLEDGMENT

The author wishes to thank Prof. R. S. Michalski for his help in the preparation of this paper.

REFERENCES

[1] T. M. Cover, "Capacity problems for linear machines," in *Pattern Recognition*, L. Kanal, Ed. Washington, DC: Thompson Book Co., 1968.

[2] ——, "Classification and generalization capabilities of linear threshold logic units," Rome Air Develop. Center Tech. Documentary, Rep. RADC-TDR-64-32, Feb. 1964.

[3] N. T. Zagoruiko, *Metody raspoznavaniia i ikh primenenie*, Sovetskoe Radio, Moskva, 1972.

[4] R. S. Michalski and B. H. McCormick, "Interval generalization of switching theory," Univ. of Illinois Dep. Comput. Sci. Rep. 442, May 1971.

[5] R. S. Michalski, "A geometrical model for the synthesis of interval covers," Univ. of Illinois Dep. Comput. Sci. Rep. 461, June 1971.

[6] N. J. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965.

[7] T. M. Cover, "Geometrical and statistical properties of linear threshold devices," Stanford Electron. Lab. Tech. Rep. 6107-1, May 1964.

[8] R. J. Brown, "Adaptive multiple-output threshold systems and their storage capacities," Stanford Electron. Lab., Stanford Univ., Tech. Rep. 6671-1, June 1964.

[9] D. E. Knuth, *The Art of Computer Programming*, vol. 2. Reading, MA: Addison-Wesley, 1969, pp. 422–444.

[10] R. S. Michalski, "Variable-valued logic: System VL_1," Presented at the 1974 Int. Symp. Multiple-Valued Logic, West Virginia Univ., Morgantown, WV, May 29–31, 1974.

[11] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application to pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, pp. 326–334, June 1965.

[12] ——, "The number of inducible orderings of points in d-space," *SIAM J. Appl. Math.*, vol. 15, pp. 434–439, Mar. 1967.

[illegible] [illegible]:

J.D. Ullman — [illegible]

[illegible] — Combinatorial Optimization