Approximating Discrete Probability Distributions with Dependence Trees

C. K. CHOW, SENIOR MEMBER, IEEE, AND C. N. LIU, MEMBER, IEEE

***Abstract*—A method is presented to approximate optimally an n-dimensional discrete probability distribution by a product of second-order distributions, or the distribution of the first-order tree dependence. The problem is to find an optimum set of $n - 1$ first order dependence relationship among the n variables. It is shown that the procedure derived in this paper yields an approximation of a minimum difference in information. It is further shown that when this procedure is applied to empirical observations from an unknown distribution of tree dependence, the procedure is the maximum-likelihood estimate of the distribution.**

I. Introduction

IN DESIGNING many information systems, such as communication, pattern-recognition, and learning systems, a central problem is to estimate the underlying n-dimensional probability distributions from a finite number of samples and to store the distributions in a certain limited amount of machine memory. Limitations on allowable equipment complexity and available samples often require the distributions to be approximated by the use of some simplifying assumptions, such as the statistical independence or the normality of the random variables under consideration. The performance of these systems is, to a very great extent, determined by the approximations employed. The aim of this paper is to apply a notion of tree dependence to approximate probability distributions. The present concern is with n-dimensional discrete distributions.

Lewis[1] and Brown[2] considered the problem of approximating an nth-order binary distribution by a product of several of its component distributions of lower order. Lewis showed that the product approximation, under suitably restricted conditions, has the property of minimum information.[1] However, the problem of selecting a set of component distributions of a given complexity to compose the best approximation remains unsolved. A method is developed in this paper to best approximate an nth-order distribution by a product of $n - 1$ second-order component distributions.

In many applications, the probability distribution function is not explicitly given, and it is usually necessary to construct a distribution function from the samples. The optimum approximation procedure is extended to empirical observations. It is shown that our procedure maximizes the likelihood function, and, therefore, it is a maximum-likelihood estimator of the distribution of tree dependence.

Manuscript received May 12, 1967; revised November 8, 1967. A preliminary version of this paper was presented at the First Annual Princeton Conference on Information Sciences and Systems, Princeton, N. J., March 1967. An abstract appeared in the Proceedings of the Conference.

The authors are with the Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, N. Y. 10598

II. A Class of Product Approximations

Let $P(\mathbf{x})$ be a joint probability distribution of n discrete variables $x_1, x_2, \cdots, x_n$, $\mathbf{x}$ denoting the n vector $(x_1, x_2, \cdots, x_n)$. A product approximation of $P(\mathbf{x})$ is defined to be a product of several of its component distributions of lower order in such a way that the product is a probability extension of these distributions of lower order.[1] Any product approximation, by definition, is itself a valid probability distribution.

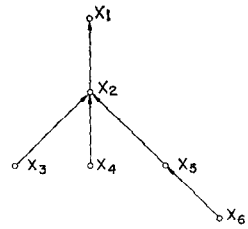
We shall consider the class of product approximations in which only the second-order distributions are used. There are $n(n - 1)/2$ second-order approximations, of which at most $n - 1$ can be used in the product approximant. In other words, the probability distributions that are permissible as approximations are of the following form[3]:

$$P_t(\mathbf{x}) = \prod_{i=1}^{n} P(x_{m_i} \mid x_{m_{j(i)}}), \qquad 0 \leq j(i) < i \qquad (1)$$

where $(m_1, \cdots, m_n)$ is an unknown permutation of integers $1, 2, \cdots, n$, and $P(x_i \mid x_0)$ is by definition equal to $P(x_i)$. Each variable in the above expansion may be conditioned upon at most one of the variables. A probability distribution that can be represented as in (1) is called a probability distribution of first-order tree dependence. The pair consisting of the set $\mathbf{x} = \{x_i \mid i = 1, 2, \cdots, n\}$ and the mapping $j(i)$ with $0 \leq j(i) < i$ is called the dependence tree of the distribution.

The following discussions are confined mainly to the first-order dependence; hence, the adjective "first-order" will be omitted whenever tolerable. For simplicity, in the following sections we will represent $(m_1, m_2, \cdots, m_n)$, the permutation of integers $1, 2, \cdots, n$, by the subscripts only; for example, x_{m_i} would be represented by x_i.

To depict the dependence relations graphically, the variable x_i will be represented by a point on the plane, and if x_l and x_m are two variables such that $m = j(l)$, they will be joined by a line with an arrowhead pointing from x_l to x_m. Whenever $j(l) = 0$, x_l will not have a line pointing away from x_l. If $j(i) = 0$ for exactly one variable, then the dependence tree is connected and has $n - 1$ branches; hence, it is a tree in the graph-theoretical sense. Otherwise, the dependence tree is a subgraph of a tree. Fig. 1 shows an example of a dependence tree.



$$P(X) = P(X_1)\,P(X_2|X_1)\,P(X_3|X_2)\,P(X_4|X_2)\,P(X_5|X_2)\,P(X_6|X_5)$$

Fig. 1. Example of a dependence tree.

III. Optimum Approximations

A probability distribution, like any other function, can be approximated by a number of different procedures. In this paper, we consider the problem of best approximating an nth-order distribution by a product of $n - 1$ second-order distributions. It is of considerable importance, both for theory as well as for practical application, to accomplish as much as possible with distributions of a fixed and low order. It goes without saying that in order to achieve increasing accuracy in the approximations, the approximants will, in general, have to be of increasingly high order.

In order to discuss the goodness of approximation, the notion of closeness of approximation must be first defined. Let $P(\mathbf{x})$ and $P_a(\mathbf{x})$ be two probability distributions of n discrete variables $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. It is well known[8] that the quantity

$$I(P, P_a) = \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{P_a(\mathbf{x})} \tag{2}$$

has the property that

$$I(P, P_a) \geq 0 \tag{3}$$

with equality sign if and only if $P(\mathbf{x}) \equiv P_a(\mathbf{x})$ for all $\mathbf{x}$. Lewis[1] defined $I(P, P_a)$ as the measure of closeness in approximating P by P_a on the basis that $I(P, P_a)$ can be interpreted as the difference of the information contained in $P(\mathbf{x})$ and that contained in $P_a(\mathbf{x})$ about $P(\mathbf{x})$. The measure is always positive if two distributions are different, and is zero if they are identical. Lewis further found that the closeness measure is particularly simple when applied to product expansions, and used the measure for comparison of two or more proposed approximations.

The measure defined in (2) will be used as a criterion in developing a procedure of approximating an nth-order distribution by a distribution of tree dependence. The problem can be stated as follows.

A Minimization Problem

Given an nth-order probability distribution $P(x_1, x_2, \cdots, x_n)$, x_i being discrete, we wish to find a distribution of tree dependence $P_\tau(x_1, x_2, \cdots, x_n)$ such that $I(P, P_\tau) \leq I(P, P_t)$ for all $t \in T_n$ where T_n is the set of all possible first-order dependence trees. The solution τ is called the optimal first-order dependence tree.

Since there are n^{n-2} trees with n vertices,[6] the number of dependence trees in T_n for any moderate value of n is so enormous as to exclude any approach of exhaustive search. To describe our solution to this optimization problem, we shall make the following definitions.

Definition 1: The mutual information $I(x_i, x_j)$ between two variables x_i and x_j is given by

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \left(\frac{P(x_i, x_j)}{P(x_i)P(x_j)} \right).$$

This is the usual definition of mutual information. It is well known that $I(x_i, x_j)$ is non-negative.

In the graphical representation of dependence relations, to every branch of the dependence tree we assign a branch weight $I(x_i, x_{j(i)})$. Given a dependence tree t, the sum of all branch weights is a useful quantity.

Definition 2: A maximum-weight dependence tree is a dependence tree t such that for all t' in T_n

$$\sum_{i=1}^{n} I(x_i, x_{j(i)}) \geq \sum_{i=1}^{n} I(x_i, x_{j'(i)}).$$

The first result can now be stated as follows.

A probability distribution of tree dependence $P_t(\mathbf{x})$ is an optimum approximation to $P(\mathbf{x})$ if and only if its dependence tree t has maximum weight.

To show this, we have from (2)

$$\begin{aligned} I(P, P_t) &= -\sum_{\mathbf{x}} P(\mathbf{x}) \sum_{i=1}^{n} \log P(x_i \mid x_{j(i)}) \\ &\quad + \sum_{\mathbf{x}} P(\mathbf{x}) \log P(\mathbf{x}) \\ &= -\sum_{\mathbf{x}} P(\mathbf{x}) \sum_{\substack{i=1 \\ j(i)\neq 0}}^{n} \log \frac{P(x_i, x_{j(i)})}{P(x_i)P(x_{j(i)})} \\ &\quad - \sum_{\mathbf{x}} P(\mathbf{x}) \sum_{i=1}^{n} \log P(x_i) \\ &\quad + \sum_{\mathbf{x}} P(\mathbf{x}) \log P(\mathbf{x}). \end{aligned} \tag{4}$$

Since $P(x_i)$ and $P(x_i, x_{j(i)})$ are components of $P(\mathbf{x})$,

$$-\sum_{\mathbf{x}} P(\mathbf{x}) \log P(x_i) = -\sum_{x_i} P(x_i) \log P(x_i)$$

which is denoted by $H(x_i)$ and

$$\begin{aligned} \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(x_i, x_{j(i)})}{P(x_i)P(x_{j(i)})} &= \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)}) \log \frac{P(x_i, x_{j(i)})}{P(x_i)P(x_{j(i)})} \\ &= I(x_i, x_{j(i)}). \end{aligned}$$

Thus, (4) becomes

$$I(P, P_t) = -\sum_{i=1}^{n} I(x_i, x_{j(i)}) + \sum_{i=1}^{n} H(x_i) - H(\mathbf{x}).$$

Since $H(\mathbf{x})$ and $H(x_i)$ for all i are independent of the dependence tree and $I(P, P_t)$ is non-negative, minimizing the closeness measure $I(P, P_t)$ is equivalent to maximizing the total branch weight

$$\sum_{i=1}^{n} I(x_i, x_{j(i)}).$$

By virtue of this result, our minimization problem can be solved without exhaustively considering all possible expansions. Furthermore, the solution is achieved without requiring any knowledge of the actual distributions of higher order other than what is necessary to evaluate the mutual information between pairs of variables. The second-order component distributions suffice for this purpose.

A direct solution is possible because the problem of finding the optimal first-order dependence tree is transformed to that of maximizing the total branch weight of a dependence tree. Since the branch weights are additive, the maximum-weight dependence tree can thus be constructed branch by branch. A procedure to best approximate an nth-order distribution by a second-order product expansion is described in Section IV.

IV. An Optimization Procedure

We shall make use of the result proved in Section III to define an optimization procedure. Our problem is to construct a dependence tree of maximum weight. We can use a simple algorithm developed by Kruskal for the construction of trees of minimum total length.[4] To choose a tree of maximum total branch weight, we first index the $n(n-1)/2$ branches according to decreasing weights, so that the weight of b_i is greater than or equal to the weight of b_j whenever $i < j$. We then start by selecting b_1 and b_2, and add b_3 if b_3 does not form a cycle with b_1 and b_2. In general, we continue to consider branches of successively higher indices, selecting a branch whenever it does not form a cycle with the set previously selected, and rejecting it otherwise. This procedure produces a unique solution if the branch weights are all different. If several weights are equal, multiple solutions are possible; however, these solutions all have the same maximum weight.

In order to provide a more detailed description of this procedure, we call attention to the flow diagram describing the computational algorithm in Fig. 2. In this figure, the input to the program is a set of samples from the distribution that is being approximated. On the basis of these samples, all $n(n-1)/2$ pairwise mutual information measures $I(x_i, x_j)$, $i = 1, 2, 3, \cdots, n-1$, $j = 2, 3, \cdots, n$, and $i < j$ are first computed. If $P(\mathbf{x})$ is explicitly given, then $I(x_i, x_j)$ are directly evaluated, and no sample is needed. The successive steps in selecting the branches are obvious from the flow diagram. When all branches are determined, $P_t(\mathbf{x})$ can be readily formed.

To understand the approximation method described in this section better, let us consider a simple example. In this example, it is desired to find the optimum tree approximation of a fourth-order binary distribution.

Example: Consider the probability distribution listed in Table I. For each pair of variables (x_i, x_j), we calculate the mutual information $I(x_i, x_j)$. These quantities are

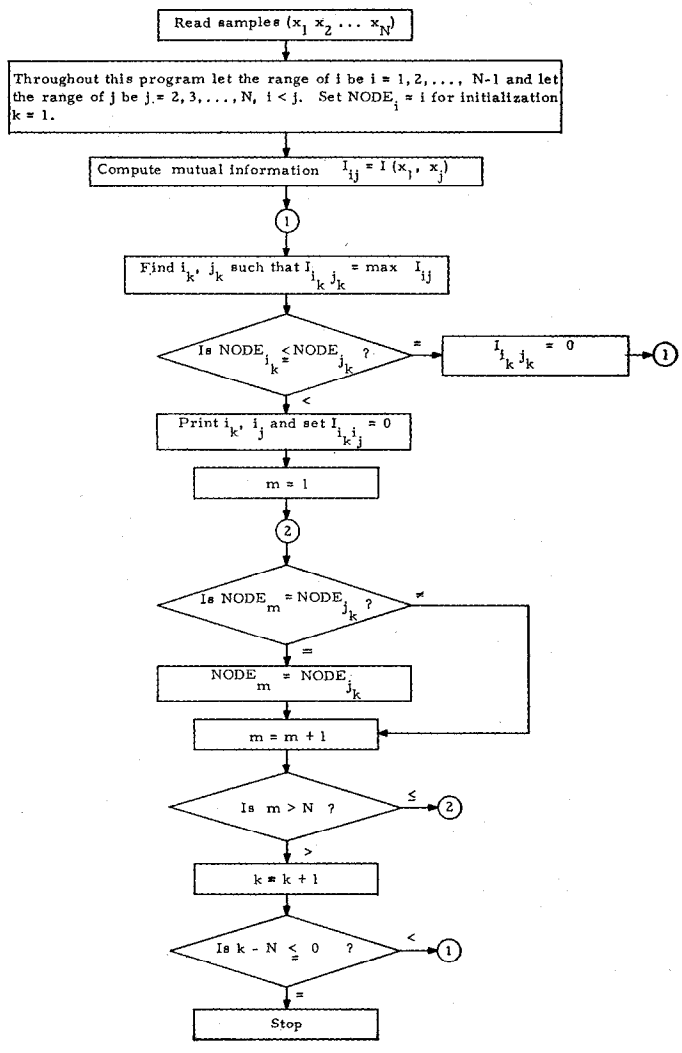


Fig. 2. Flow diagram of the optimization procedure.

given below and on the branches joining the nodes in Fig. 3. Natural logarithms are used in the computation.

$$I(x_1, x_2) = 0.079$$
$$I(x_1, x_3) = 0.00005$$
$$I(x_1, x_4) = 0.0051$$
$$I(x_2, x_3) = 0.189$$
$$I(x_2, x_4) = 0.0051$$
$$I(x_3, x_4) = 0.0051.$$

Since $I(x_2, x_3)$ and $I(x_1, x_2)$ are the two largest quantities, (x_2, x_3) and (x_1, x_2) constitute the first two branches of the optimum dependence tree. To select the next branch, we note that $I(x_1, x_4) = I(x_2, x_4) = I(x_3, x_4)$. Our program usually would pick arbitrarily any one of these three branches and proceed to the next branch. However, since this is the last branch to be selected in this example, we accept all three alternatives and list their corresponding probabilities in Table II.

For comparison purposes, the approximant with the assumption of statistical independence is also listed in

TABLE I
A BINARY PROBABILITY DISTRIBUTION

$x_1\ x_2\ x_3\ x_4$	$P(x_1\ x_2\ x_3\ x_4)$	$p(x_1)\ p(x_2)\ p(x_3)p(x_4)$
0 0 0 0	0.100	0.046
0 0 0 1	0.100	0.046
0 0 1 0	0.050	0.056
0 0 1 1	0.050	0.056
0 1 0 0	0.000	0.056
0 1 0 1	0.000	0.056
0 1 1 0	0.100	0.068
0 1 1 1	0.050	0.068
1 0 0 0	0.050	0.056
1 0 0 1	0.100	0.056
1 0 1 0	0.000	0.068
1 0 1 1	0.000	0.068
1 1 0 0	0.050	0.068
1 1 0 1	0.050	0.068
1 1 1 0	0.150	0.083
1 1 1 1	0.150	0.083

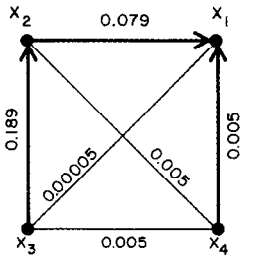


Fig. 3. Selection of an optimization dependence tree.

TABLE II
OPTIMAL TREE APPROXIMATION

$x_1\ x_2\ x_3 x_4$	$p(x_1)\,p(x_2/x_1)\cdot p(x_3/x_2)\,p(x_4/x_1)$	$p(x_1)\,p(x_2/x_1)\cdot p(x_3/x_2)\,p(x_4/x_2)$	$p(x_1)\ p(x_2/x_1)\ p(x_3/x_2)\ P(x_4/x_3)$
0 0 0 0	0.130	0.104	0.104
0 0 0 1	0.104	0.130	0.130
0 0 1 0	0.037	0.030	0.037
0 0 1 1	0.030	0.037	0.030
0 1 0 0	0.015	0.015	0.012
0 1 0 1	0.012	0.012	0.015
0 1 1 0	0.068	0.068	0.068
0 1 1 1	0.054	0.055	0.055
1 0 0 0	0.053	0.052	0.052
1 0 0 1	0.064	0.065	0.065
1 0 1 0	0.015	0.015	0.018
1 0 1 1	0.018	0.019	0.015
1 1 0 0	0.033	0.040	0.032
1 1 0 1	0.040	0.032	0.040
1 1 1 0	0.149	0.182	0.182
1 1 1 1	0.178	0.146	0.146

Table I. A comparison of the figures on Tables I and II shows that the optimum tree approximants are closer to the true distribution. The closeness measure of approximation $I(P, P_a)$ for any of the three optimum distributions of tree dependence is 0.094; that for the independent distribution is 0.364.

We have illustrated here only a simple problem, but the computational advantage of our technique becomes more and more prominent as the combinatorial feature becomes magnified for larger values of n.

V. ESTIMATION

In applications, the probability distribution is frequently not explicitly given and only samples are available. It is necessary to construct a distribution from the samples. This situation is typical in most pattern-recognition problems. To achieve a second-order product approximation, the dependence tree, in addition to the parameters, must be estimated. The problem of estimating from samples the values of parameters has been extensively treated by statisticians, and many methods are available for such estimation. However, the problem of constructing dependence trees makes a new method necessary.

A method is developed in this paper to construct an optimal dependence tree from samples. Two approaches are possible. The one that seems more natural in the present context is to extend the optimization procedure to empirical observations; the other is to apply the principle of maximum likelihood (see Appendix). It is significant that both approaches lead to the same estimation procedure.

Let $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^s$ be s independent samples of a finite discrete variate $\mathbf{x}$. $\mathbf{x}$ is the vector $(x_1, x_2, \cdots, x_n)$ and $\mathbf{x}^k = (x_1^k, x_2^k, \cdots, x_n^k)$. The estimation procedure is as follows.

1) Compute for all pairs of variables x_i and x_j the sample joint frequencies $f(x_i, x_j)$ as

$$f_{uv}(i, j) = \frac{n_{uv}(i, j)}{\sum_{uv} n_{uv}(i, j)} \tag{5}$$

and

$$f_u(i) = \sum_v f_{uv}(i, j)$$

where $f_{uv}(i, j)$ and $f_u(i)$ denote $f(x_i = u, x_j = v)$ and $f(x_i = u)$, respectively, and $n_{uv}(i, j)$ is the number of samples such that their ith and jth components assume the values of u and v, respectively. It is well known that $f_{uv}(i, j)$ is a maximum-likelihood estimator for the probability $P(x_i = u, x_j = v)$.

2) Compute for all i and j the sample mutual information $\hat{I}(x_i, x_j)$ as

$$\hat{I}(x_i, x_j) = \sum_{u,v} f_{uv}(i, j) \log \frac{f_{uv}(i, j)}{f_u(i) f_v(j)}.$$

3) Take $\hat{I}(x_i, x_j)$ as $I(x_i, x_j)$ and use the optimization procedure to obtain a tree τ such that the tree sum of mutual information

$$\sum_{i=1}^{n} \hat{I}(x_i, x_{j(i)})$$

is maximized.

The procedure minimizes the sample value of the closeness measure. Furthermore, it can be shown (see Appendix) that the procedure also maximizes the likelihood function and is a maximum-likelihood estimator for the dependence tree. The consistency property of maximum-likelihood estimates also holds for our procedure. In consequence, if the underlying distribution is one of tree dependence, then the tree rendered by the present precedure converges with probability one to the true tree of dependence.

In the following application to a pattern-recognition problem, the probability distributions are estimated by this procedure.

VI. Application to Pattern Recognition

Pattern recognition can be considered as a statistical decision problem. Within the framework of a statistical decision approach, the structure of optimum recognition systems depends upon a set of conditional probability distributions.

Let c be the number of pattern classes, and let a_i denote the ith class. Let $\boldsymbol{p} = (p_1, p_2, \cdots, p_c)$ be the a priori distribution of the classes. An unknown pattern represented by a measurement vector $\mathbf{x} = (x_1, x_2 \cdots, x_n)$ is decided to be a sample of class a_k if

$$p_k P(\mathbf{x} \mid a_k) \geq p_i P(\mathbf{x} \mid a_i)$$

for all i.

Since the information about the patterns is generally contained in a set of samples, the central problem here is to estimate or to approximate the unknown conditional probability distributions $P(\mathbf{x} \mid a_i)$. It is reasonable to assume that optimum approximations of the conditional probabilities $P(\mathbf{x} \mid a_i)$ would lead to effective recognition.

The problem of recognizing hand-printed numerals was investigated. Approximately 19 000 numerals produced in the course of routine operations by four inventory clerks in a department store were scanned by a CRT scanner. Samples of scanned numerals are shown in Fig. 4. Ninety-six binary measurements[7] were used to represent the numerals. Samples were divided into two subsets; the first subset consisting of 6947 samples was employed as design data, and the remaining 12 000 samples were used for testing. The tree approximation program was used to derive 10 optimum dependence trees, one for each class of numerals. Fig. 5 shows the first part of the dependence tree derived for numeral 4. Recognition based on a procedure discussed in Chow[3] was tried on the test data set. Results are depicted by the error versus rejection curve in Fig. 6.

For comparison purposes, results from assuming the independence of measurements are also plotted in Fig. 6. A reduction in error rate by a factor of 2 was realized by the tree approximations. This reduction is considered

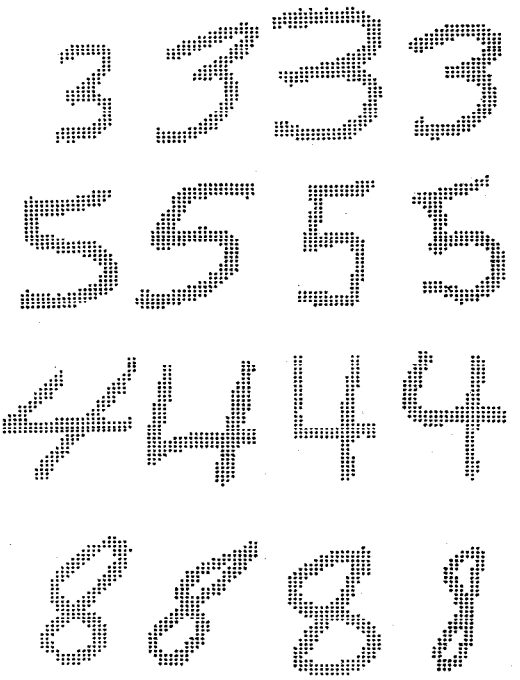


Fig. 4. Samples of scanned numerals.

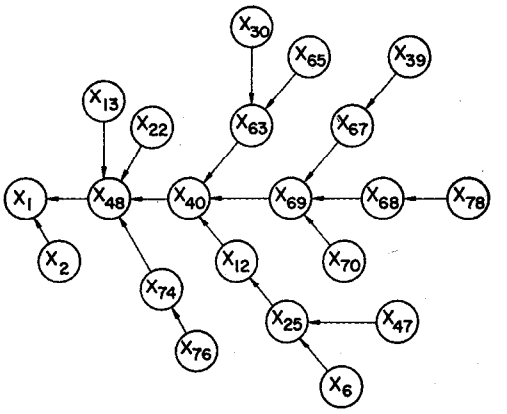


Fig. 5. First part of dependence tree for numeral 4.

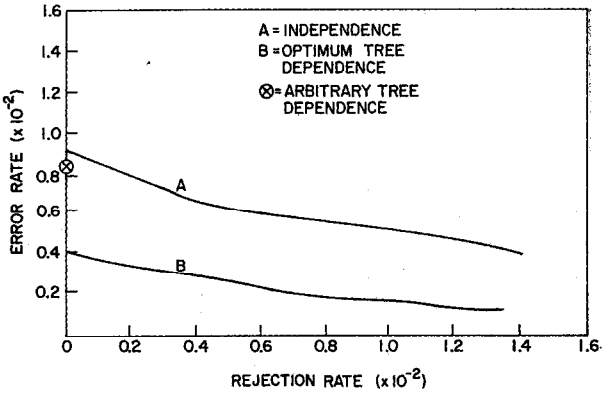


Fig. 6. Error versus rejection curves.

significant, although it may not be the best achievable with a second-order product approximation, since the present approximation criterion is not that of minimizing the recognition-error rate. Recently, some progress has been reported[5] toward minimization of error rate; however, there is yet no general solution.

In order to examine the effectiveness of the optimum dependence tree as compared with an arbitrary dependence tree, the following recognition experiment was simulated. Each measurement x_i, $i = 2, 3, \cdots, 96$ was assumed to depend upon its predecessor x_{i-1}; $P(x_1 \mid x_0)$ by definition is equal to $P(x_1)$. The resulting decision function misrecognized 0.829 percent of the test sample set. Although this percentage is somewhat lower than that (0.901 percent) obtained with the assumption of measurement independence, it is still higher by about a factor of 2 than that (0.417 percent) obtained with the optimum dependence trees.

VII. Conclusions

The viewpoint and mathematical model described in an earlier paper for the design of pattern recognition networks are applied here to the problem of approximating an nth-order probability distribution by a particular class of distributions in which each variable is conditioned upon, at most, one other variable. When an information measure is used as the criterion of goodness of approximation, necessary and sufficient conditions for the optimum approximation of a given nth-order probability distribution are derived. Consequently, an efficient computational algorithm is obtained.

It is further shown that when only samples drawn from the nth-order distribution are available, the optimization procedure can be extended to construct the dependence tree as well as to estimate the parameters. In fact, the procedure leads to a maximum-likelihood estimator of the dependence tree.

In addition to the theoretical studies, some experiments are carried out to investigate the effectiveness of the present method as applied to the recognition of handwritten numerals. It is found that significant improvement of recognition performance may be realized with the present procedure.

Appendix

A Maximum-Likelihood Estimator

The estimation procedure as described in Section V is the maximum-likelihood estimate of the dependence tree. A sketch of the proof is included here.

The likelihood function of s independent observations $x^1, x^2, \cdots$, and x^s from the distribution of tree dependence $P_t(\mathbf{x})$ is

$$L_t(x^1, x^2, \cdots, x^s) = \prod_{k=1}^{s} P_t(x^k) = \prod_{k=1}^{s} \prod_{i=1}^{n} P(x_{m_i}^k \mid x_{m_{j(i)}}^k) \tag{6}$$

where $(m_1, m_2, \cdots, m_n)$ is an unknown permutation of integers $1, 2, \cdots, n$; and its logarithm, after an interchange of the orders of summation, is

$$l_t(x^1, \cdots, x^s) = \log L_t(x^1, \cdots, x^s) = \sum_{i=1}^{n} \sum_{k=1}^{s} \log P(x_{m_i}^k \mid x_{m_{j(i)}}^k). \tag{7}$$

The last expression is to be maximized by selecting a tree and its associated conditional probabilities. This maximization is achieved in two steps:

$$\max_t [l_t(x^1, x^2, \cdots, x^s)] = \max_t \left[\sum_{i=1}^{n} \sum_{k=1}^{s} \log P(x_{m_i}^k \mid x_{m_{j(i)}}^k) \right] = \max_t \left\{ \sum_{i=1}^{n} \max_{P \mid t} \left[\sum_{k=1}^{s} \log P(x_{m_i}^k \mid x_{m_{j(i)}}^k) \right] \right\}. \tag{8}$$

The inner sum, for a given tree t, is maximized when the observed sample frequencies (5) are used as the estimates of the conditional probabilities. Consequently, (8), with some algebraic manipulations, becomes

$$\max_t [l_t(x^1, x^2, \cdots, x^s)] = \max_t \left[\sum_{i=1}^{n} \hat{I}(x_{m_i}, x_{m_{j(i)}}) \right] + K \tag{9}$$

where K is $\sum_{i=1}^{n} \sum_{k=1}^{s} \log P(x_{m_i}^k)$ evaluated over the sample and is independent of the tree t.

The remaining problem now is to choose the tree t such that the sum on the right-hand side of (9) is maximum. This is achieved by employing the optimization procedure described in this paper, with $\hat{I}(x_i, x_j)$ as the branch weights. The proof is thus established.

Acknowledgment

The authors wish to thank Dr. J. A. McLaughlin for his helpful discussions and encouragement. They are also grateful to L. Loh for his programming assistance.

References

[1] P. M. Lewis, "Approximating probability distributions to reduce storage requirement," *Information and Control*, vol. 2, pp. 214–225, September 1959.

[2] D. T. Brown, "A note on approximations to discrete probability distributions," *Information and Control*, vol. 2, pp. 386–392, December 1959.

[3] C. K. Chow, "A class of nonlinear recognition procedures," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-2, pp. 101–109, December 1966.

[4] J. B. Kruskal, Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Am. Math. Soc.*, vol. 7, pp. 48–50, 1956.

[5] C. K. Chow and C. N. Liu, "An approach to structure adaptation in pattern recognition," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-2, pp. 73–80, December 1966.

[6] H. M. Trent, "A note on the enumeration and listings of all possible trees in a connected linear graph," *Proc. Nat'l Acad. Sciences*, vol. 40, pp. 1004–1007, 1954.

[7] R. Bakis, N. Herbst, and G. Nagy, "An experimental study of machine recognition of hand-printed numerals," *IEEE Trans. Systems Science and Cybernetics* (to be published).

[8] S. Kullback and R. A. Leiber, "On information and sufficiency," *Ann. Math. Statistics*, vol. 22, pp. 79–86, 1951.