Comments on Approximating Discrete Probability Distributions with Dependence Trees

S. K. M. WONG AND F. C. S. POON

***Abstract*—Chow and Liu introduced the notion of tree dependence to approximate a *k*th order probability distribution. More recently, Wong and Wang proposed a different product approximation. The aim of this paper is to show that the tree dependence approximation suggested by Chow and Liu can be derived by minimizing an upper bound of the Bayes error rate under certain assumptions. It is also shown that the method proposed by Wong and Wang does not necessarily lead to fewer misclassifications because it is a special case of such a minimization procedure.**

***Index Terms*—Bayes error rate, classification, entropy, information theory, mutual information, pattern recognition, probability distribution, tree dependence.**

I. Introduction

The problem of *classification* is one of the main concerns in the design of intelligent information systems such as pattern recognition, inductive learning, and expert systems. In many of these applications, the essential task is to estimate the underlying k-dimensional probability distributions from a finite set of samples. Because of the curse of dimensionality, the probability distribution function is often approximated by some simplifying assumptions, such as statistical independence. The independence approximation is simple but may be unrealistic in certain applications. It was suggested by Lewis [1] that the optimal product approximation can be obtained by minimizing a divergence measure between the true and approximate distributions. Some years ago Chow and Liu [2] introduced the notion of *tree* dependence to approximate a kth-order probability distribution by a product of $k - 1$ second-order component distributions. One can then reduce the problem to finding a dependence tree with maximum total branch weight of mutual information [2], [3].

It was mentioned in [2] that the tree selection criterion is not that of minimizing the recognition-error rate (Bayes error rate). More recently, Wong and Wang suggested another product approximation by minimizing an *upper bound* of the Bayes error rate. (This method is referred to as "Error Probability Minimax" in [4], [5].)

The aim of this correspondence is to show that the tree dependence approximation proposed by Chow and Liu can in fact be derived by minimizing an upper bound of the Bayes error rate under certain assumptions. Moreover, we show that the method suggested by Wong and Wang is a special case of such a minimization procedure.

II. Tree Dependence Approximation Based on Minimization of Bayes Error Rate

Before we discuss the approach proposed by Wong and Wang, we first show that the results of Chow and Liu can be obtained from a minimization procedure.

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)$ denote a k-dimensional random vector. The component X_i of $\boldsymbol{X}$ represents the ith discrete-valued feature. Let $\boldsymbol{W}$ be a random variable whose values are used to label the classes. Let $P(x, \omega)$ be the true joint probability distribution for $\boldsymbol{X} = \boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ and $\boldsymbol{W} = \omega$, where $\boldsymbol{x}$ is a value of the random vector $\boldsymbol{X}$. The probability distributions that are permissible as approximations can be written as

$$\hat{P}(\boldsymbol{x}, \omega) = \prod_{i=1}^{n} P(x_{m_i} | x_{m_{j(i)}}, \omega), \qquad 0 \le j(i) < i \qquad (1)$$

where $(m_1, \cdots, m_n)$ is an unknown permutation of the integers $1, 2, \cdots, n$, $P(x_{m_i} | x_{m_{j(i)}}, \omega)$ the joint probability of x_{m_i} and ω conditioned on the variable $x_{m_{j(i)}}$, and $P(x_i | x_0, \omega)$ by definition equal to $P(x_i, \omega)$. For notation convenience, we will drop the subscript m and denote, for example, x_{m_i} by x_i in subsequent discussions.

Let σ_e denote the Bayes error rate. It was proved by Hellman and Raviv [6] that

$$\sigma_e \le \tfrac{1}{2} H(\omega | \boldsymbol{X}), \qquad (2)$$

where the entropy function $H(\omega | \boldsymbol{X})$ is defined by

$$H(\omega | \boldsymbol{X}) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \sum_{\omega} P(\omega | \boldsymbol{x}) \log P(\omega | \boldsymbol{x}).$$

The function $H(\omega | \boldsymbol{X})$ can be rewritten as

$$H(\omega | \boldsymbol{X}) = H(\omega) - H(\boldsymbol{X}) - \sum_{\omega} P(\omega) \sum_{\boldsymbol{x}} P(\boldsymbol{x} | \omega) \log P(\boldsymbol{x} | \omega) \qquad (3)$$

where

$$H(\omega) = -\sum_{\omega} P(\omega) \log P(\omega),$$
$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x}).$$

In terms of the second-order approximation defined by (1) for each individual class ω

$$\hat{P}(\boldsymbol{x} | \omega) = \frac{1}{P(\omega)} \prod_{i=1}^{n} P(x_i | x_{j(i)}, \omega), \qquad 0 \le j(i) < i, \qquad (4)$$

we obtain from (3),

$$\begin{aligned} \hat{H}(\omega | \boldsymbol{X}) &= H(\omega) - H(\boldsymbol{X}) - \sum_{\omega} P(\omega) \sum_{\boldsymbol{x}} \hat{P}(\boldsymbol{x} | \omega) \log \hat{P}(\boldsymbol{x} | \omega) \\ &= H(\omega) - H(\boldsymbol{X}) - \sum_{\omega} P(\omega) \sum_{i=1}^{n} I_\omega(X_i, X_{j(i)}) \\ &\quad - \sum_{\omega} P(\omega) \sum_{i=1}^{n} H_\omega(X_i) \qquad (5) \end{aligned}$$

where

$$I_\omega(X_i, X_{j(i)}) = \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)} | \omega) \log \frac{P(x_i, x_{j(i)} | \omega)}{P(x_i | \omega) \log P(x_{j(i)} | \omega)},$$
$$H_\omega(X_i) = -\sum_{x_i} P(x_i | \omega) \log P(x_i | \omega).$$

If we assume that the $H(\boldsymbol{X})$ is independent of the dependence tree chosen for each individual class, by minimizing $\hat{H}(\omega | \boldsymbol{X})$ defined above it follows:

$$\min \hat{H}(\omega | \boldsymbol{X}) = \max \sum_{\omega} \sum_{i=1}^{n} I_\omega(X_i, X_{j(i)}), \qquad (6)$$

which is the result obtained by Chow and Liu. Kruskal's algorithm [3] can be easily applied to finding a tree with maximum total

Manuscript received February 3, 1987; revised December 23, 1987. Recommended for acceptance by A. K. Jain.

The authors are with the Department of Computer Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada.

IEEE Log Number 8824647.

0162-8828/89/0300-0333$01.00 © 1989 IEEE

branch weight,

$$B_\omega = \sum_{i=1}^{n} I_\omega(X_i, X_{j(i)}), \tag{7}$$

for each individual class ω.

On the other hand, one may assume that the probability distributions for *all* the classes can be approximated by the *same* dependence tree as suggested by Wong and Wang [3], [4]. In this case, for $0 \le j(i) < i$, the *a priori* probability distribution $P(\boldsymbol{x})$ can be written as:

$$\hat{P}(\boldsymbol{x}) = \prod_{i=1}^{n} P(x_i | x_{j(i)}). \tag{8}$$

By substituting the approximate distributions defined by (1) and (8) into (3), one immediately obtains the following result of Wong and Wang:

$$\min \hat{H}(\omega | \boldsymbol{X}) = \max \sum_{i=1}^{n} \left(\sum_{\omega} P(\omega)\, I_\omega(X_i, X_{j(i)}) - I(X_i, X_{j(i)}) \right) \tag{9}$$

where

$$I(X_i, X_{j(i)}) = \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)}) \log \frac{P(x_i, x_{j(i)})}{P(x_i)\, P(x_{j(i)})}.$$

The important point is that Chow and Liu's method uses one tree structure for *each* individual class. In contrast, by adopting the same minimization procedure the result of Wong and Wang is obtained by using one tree structure for *all* classes.

III. Experimental Results

Before presenting our experimental results, we will first demonstrate the restrictiveness of 1-*tree* method (Wong and Wang) in comparison with the 2-*tree* method (Chow and Liu) by the following example.

Example 1: Consider a sample with three features and two pattern classes ($W =$ "+" and $W =$ "−"). Each feature has a value of 0 or 1. The probability distribution within each class is shown in Table I and both $P(+)$ and $P(-)$ are equal to 0.5.

Based on the *exact* probability distributions in Table I, the classification for each feature vector $\boldsymbol{x}$ is determined by using the Bayes decision rule

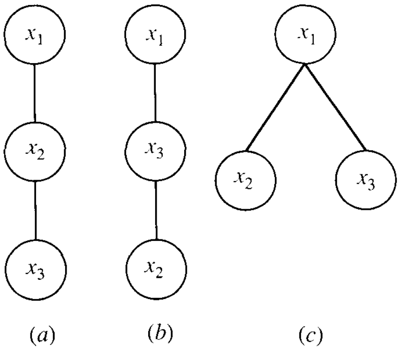
Decide +, if $P(\boldsymbol{x}|+)\,P(+) > P(\boldsymbol{x}|-)\,P(-)$

or

Decide −, if $P(\boldsymbol{x}|-)\,P(-) > P(\boldsymbol{x}|+)\,P(+)$.

The classification results are listed in the last column of the Table I.

There are three possible tree structures in this example:



(*a*) (*b*) (*c*)

TABLE I

Probability Distributions with Two Classes

Feature vector $\mathbf{x} = (x_1, x_2, x_3)$	$P(\mathbf{x}\|+)$	$P(\mathbf{x}\|-)$	Classification
0 0 0	0.072	0.0728	−
0 0 1	0.168	0.1352	+
0 1 0	0.072	0.1092	−
0 1 1	0.288	0.2028	+
1 0 0	0.036	0.0348	+
1 0 1	0.084	0.1044	−
1 1 0	0.056	0.0852	−
1 1 1	0.224	0.2556	−

1) 2-tree Method (Chow and Liu): By (6), *tree* (*a*) is selected from class "+" and *tree* (*c*) for class "−". Using the Bayes decision rule and the approximate probability distributions

$$\hat{P}(\boldsymbol{x}|+) = P(x_1|+)\,P(x_2|x_1, +)\,P(x_3|x_2, +),$$

$$\hat{P}(\boldsymbol{x}|-) = P(x_1|-)\,P(x_2|x_1, -)\,P(x_3|x_1, -),$$

one obtains the classification results shown in column 2 of Table II. By comparing these results to those of the original classification, no misclassification error is observed

2) 1-tree Method (Wong and Wang): Based on (9), *tree* (*b*) is selected to approximate both $P(\boldsymbol{x}|+)$ and $P(\boldsymbol{x}|-)$, namely,

$$\hat{P}(\boldsymbol{x}|+) = P(x_1|+)\,P(x_3|x_1, +)\,P(x_2|x_3, +),$$

$$\hat{P}(\boldsymbol{x}|-) = P(x_1|-)\,P(x_3|x_1, -)\,P(x_2|x_3, -).$$

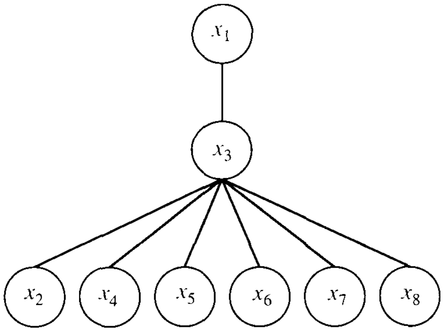
The classification results obtained by using these approximate distributions are shown in column 3 of Table II. Note that there are two misclassifications in this case.

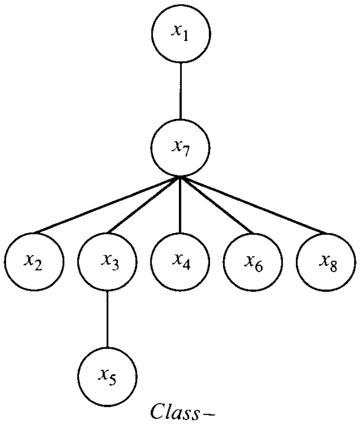
The results of this example indicate that Wong and Wang's method may lead to a higher number of misclassifications than Chow and Liu's method. □

We performed nine experiments using different probability distributions. The primary objective of these experiments is to compare the accuracy of the approximate distributions between the 2-*tree* and 1-*tree* methods. For each feature vector, we compare the original classification to the approximate one. The total number of misclassifications is used as a measure of the accuracy of the approximate under consideration.

In all our experiments, we used eight features, two classes, and various sample size up to 6172 feature vectors. In each sample we assigned a probability distribution for each class. As shown in Example 1, based on the given distributions the original classification for each feature vector was determined by the Bayes decision rule.

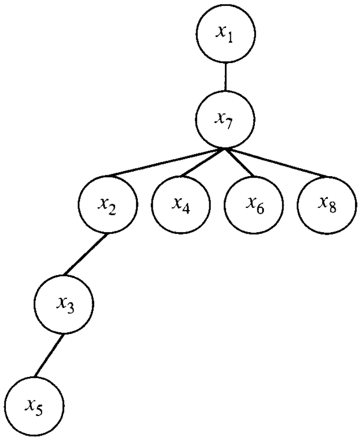
In sample 1, we used 2578 featue vectors for class "+" and 3102 for class "−". According to Chow and Liu's method, the following tree structures were selected: These two trees were then used to compute the approximate probability distributions. Based on these distributions, the classification for each feature vector was inferred from the Bayes decision rule. By comparing these results to those of the original classification, we obtained 19 misclassifications.





Class–

For the same sample, the tree structure for both classes selected by Wong and Wang's method is shown below. By applying the Bayes decision rule to the distributions of the above tree, 22 misclassifications were observed in this case.



The experimental results of other samples are summarized in Table III. In all cases except one, the 2-*tree* method performs better than the 1-*tree* method although some of the improvements are marginal.

IV. Conclusion

We have shown that the dependence tree approximation used by Chow and Liu can be derived by minimizing an upper bound of the Bayes error rate under certain assumptions. Based on our analysis, it seems that the 1-*tree* method is more restricted than the 2-*tree* method.

There is always a tradeoff between efficiency and accuracy. Obviously, Wong and Wang's method has the advantage of being computationally more efficient, especially when the number of features is very large. However, if accuracy is the predominant factor in a particular application, Chow and Liu's method is preferred.

TABLE II
Comparison of the Original and Approximate Classification

Original Classification	2–*tree* (Chow and Liu)	1–*tree* (Wong and Wang)
–	–	+
+	+	+
–	–	–
+	+	+
+	+	–
–	–	–
–	–	–
–	–	–

TABLE III
Number of Misclassifications

Sample #	Number of Misclassifications	
	2–*tree*	1–*tree*
1	19	22
2	27	29
3	9	19
4	0	3
5	6	4
6	14	18
7	4	4
8	8	8
9	0	3

References

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

[2] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462–467, May 1968.

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

[4] A. K. C. Wong and C. C. Wang, "Classification of discrete biomedical data with error probability minimax," in *Proc. SeventthInt. Conf. Cybern. Soc.*, Washington, DC, Sept. 1977, pp. 19–21.

[5] C. C. Wang and A. K. C. Wong, "Classification of discrete data with feature space transformation," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 434–437, June 1979.

[6] M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 368–372, 1970.