On the Relationship between Dependence Tree Classification Error and Bayes Error Rate

Kiran S. Balagani and
Vir V. Phoha, *Senior Member*, *IEEE*

Abstract—Wong and Poon [1] showed that Chow and Liu's tree dependence approximation can be derived by minimizing an upper bound of the Bayes error rate. Wong and Poon's result was obtained by expanding the conditional entropy $H(\omega|X)$. We derive the correct expansion of $H(\omega|X)$ and present its implication.

Index Terms—Bayes error rate, entropy, mutual information, classification, dependence tree approximation.

1 Introduction

Chow and Liu [2] introduced dependence trees to approximate nth order discrete probability distributions using a product of second-order distributions. Let $X = (X_1, \cdots, X_n)$ denote an n-dimensional discrete random feature vector. Let $W = \{\omega_1, \cdots, \omega_r\}$ be a discrete random variable whose values are the class labels. Let $P(x|\omega)$ be the conditional distribution of X given W, where $x = (x_1, \cdots, x_n)$ is a value of the feature vector X and ω is a value of W. In Chow and Liu's dependence tree approximation, the probability distribution $P(x|\omega)$ is approximated by $\hat{P}(x|\omega)$ as

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

where $(m_1, \cdots, m_n)$ is an unknown permutation of integers $1, 2, \cdots, n$, $P(x_{m_i}|x_{m_{j(i)}}, \omega)$ is a component probability in which each variable x_{m_i} is conditioned on at most one variable $x_{m_{j(i)}}$ and the component probability of the form $P(x_i|x_0, \omega)$ is by definition equal to $P(x_i|\omega)$. The unknown permutation is obtained using Kruskal's algorithm [3], which finds the spanning tree with maximum pairwise mutual information between the features. To perform dependence tree classification, the Bayes decision rule is used and the state-conditional probability distribution "$P(X|\omega)$" in the Bayes decision rule is estimated using the dependence tree approximation in (1). For notational simplicity, we will hereafter omit the subscript m of each variable and represent, for example, x_{m_i} as x_i.

Note that the dependence tree approximation, *defined as (1) in Wong and Poon's paper [1], is incorrect*. By Wong and Poon's equation, the approximation becomes an invalid probability distribution because the right-hand side of the equation does not sum to 1 over all values of x and ω.

Hellman and Raviv [4] proved that an upper bound on the Bayes error rate "σ_e" is $\frac{1}{2}H(\omega|X)$, where $H(\omega|X)$ is the conditional entropy of class ω given the n-dimensional feature vector X. Wong and Poon, in [1], extended Hellman and Raviv's result (see [6] for tighter bounds on the Bayes error rate) and showed that, under certain assumptions, Chow and Liu's dependence tree approximation procedure can be derived by minimizing the upper bound of the Bayes error rate. Wong and Poon's result comes from (5) in their paper [1], which expands the entropy function $H(\omega|X)$ and replaces $P(x|\omega)$ with probability distribution $\hat{P}(x|\omega)$ using the dependence tree approximation. The equation appeared as

$$\begin{aligned}\hat{H}(\omega|X) = H(\omega) - H(X) - \sum_{\omega} P(\omega) \sum_{i=1}^{n} I_\omega(X_i, X_{j(i)}) \\ - \sum_{\omega} P(\omega) \sum_{i=1}^{n} H_\omega(X_i),\end{aligned} \tag{2}$$

where

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

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

$$H_\omega(X_i) = -\sum_{x_i} P(x_i|\omega) \log P(x_i|\omega).$$

2 Correction and Its Implication

The correct expansion of the conditional entropy function $\hat{H}(\omega|X)$ is derived in Appendix A and is given as

$$\begin{aligned}\hat{H}(\omega|X) = H(\omega) - H(X) - \sum_{\omega} P(\omega) \sum_{i=1,\, j(i)\neq 0}^{n} I_\omega(X_i, X_{j(i)}) \\ + \sum_{\omega} P(\omega) \sum_{i=1}^{n} H_\omega(X_i).\end{aligned} \tag{3}$$

Equation (3) corrects (2) by reversing the sign of the term $\sum_\omega P(\omega) \sum_{i=1}^{n} H_\omega(X_i)$. Though this correction appears to be a minor issue, it invalidates the misleading idea purported by (2) that every component probability in the dependence tree approximation decreases the value of $\hat{H}(\omega|X)$, thereby reducing the upper bound on the Bayes error rate. The corrected equation (3) and, *more explicitly*, the expansions (9) and (10) from the derivation in Appendix A show that each component probability in the dependence tree approximation, whether in the form of $P(x_i|x_{j(i)}, \omega)$, $j(i) \neq 0$, or $P(x_i|x_0, \omega)$, *adds* $\sum_\omega P(\omega)H_\omega(X_i)$ to $\hat{H}(\omega|X)$ and does not necessarily contribute toward decreasing the upper bound on Bayes error rate. Therefore, caution is advised when selecting component probabilities for dependence tree approximation.

Below, we give two conditions to guarantee that every component probability in the dependence tree approximation decreases the value of $\hat{H}(\omega|X)$, thereby decreasing the upper bound on the Bayes error rate.

Condition 1. *In a dependence tree approximation, for each component probability of the form* $P(x_i|x_{j(i)}, \omega)$, $j(i) \neq 0$, $\sum_\omega P(\omega)I_\omega(X_i, X_{j(i)})$ *should be greater than* $\sum_\omega P(\omega)H_\omega(X_i)$.

Condition 1 follows from expansion (10) in Appendix A and concerns component probabilities of the form $P(x_i|x_{j(i)}, \omega)$, $j(i) \neq 0$, in the dependence tree approximation. We explain Condition 1 with an example. Let $X = (X_1, X_2, X_3)$ be a three-dimensional discrete random feature vector. Let $P(X_1|\omega)P(X_2|X_1, \omega)P(X_3|\omega)$ be the dependence tree approximation of $P(X|\omega)$. In this dependence tree

• *The authors are with the Computer Science, College of Engineering and Science, Louisiana Tech University, Nethken Hall, Arizona Ave., Ruston, LA 71272. E-mail: {ksb011, phoha}@latech.edu.*

Manuscript received 10 Sept. 2006; revised 6 Feb. 2007; accepted 17 May 2007; published online 29 May 2007.
Recommended for acceptance by M. Figueiredo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0658-0906.
Digital Object Identifier no. 10.1109/TPAMI.2007.1184.

approximation, there is one component probability of the form $P(x_i|x_{j(i)},\omega)$, i.e., $P(X_2|X_1,\omega)$. Expansion (9) shows that each component probability of the form $P(x_i|x_{j(i)},\omega)$ adds $\sum_\omega P(\omega)$ $H_\omega(X_i)$ to $\hat{H}(\omega|X)$. Therefore, $P(X_2|X_1,\omega)$ adds $\sum_\omega P(\omega)H_\omega(X_2)$ to $\hat{H}(\omega|X)$. However, if $\sum_\omega P(\omega)I_\omega(X_2,X_1)$ is greater than $\sum_\omega P(\omega)$ $H_\omega(X_2)$, then, from expansion (9), we see that the presence of component probability $P(X_2|X_1,\omega)$ in the dependence tree approximation decreases the value of $\hat{H}(\omega|X)$, thereby decreasing the bound on the Bayes error rate.

Condition 2. *In a dependence tree approximation, for each component probability of the form $P(x_i|x_0,\omega)$, $0<i\le n$, there must be a nonempty set l_i, $|l_i|\le n$, of component probabilities of the form $P(x_s|x_i,\omega)$, $0<s\le n$, so that $\sum_\omega P(\omega)\sum_s I_\omega(X_s,X_i)$ is greater than $\sum_\omega P(\omega)H_\omega(X_i)$.*

Condition 2 follows from (10) in Appendix A and concerns component probabilities of the form $P(x_i|x_0,\omega)$. We explain Condition 2 with an example. Let $X=(X_1,X_2,X_3,X_4)$ be a four-dimensional discrete random feature vector. Let $P(X_1|\omega)$ $P(X_2|X_1,\omega)P(X_3|X_1,\omega)P(X_4|\omega)$ be the dependence tree approximation of $P(X|\omega)$. In this dependence tree (or, more precisely, dependence forest) approximation, there are two component probabilities of the form $P(x_i|x_0,\omega)$, i.e., $P(X_1|\omega)$ and $P(X_4|\omega)$. Equation (10) shows that $P(X_1|\omega)$ and $P(X_4|\omega)$ add $\sum_\omega P(\omega)H_\omega(X_1)$ and $\sum_\omega P(\omega)H_\omega(X_4)$ to $\hat{H}(\omega|X)$. Now, consider the component probability $P(X_1|\omega)$. From Condition 2, l_1 contains all of the component probabilities conditioned on X_1, i.e., $l_1=\{P(X_2|X_1,\omega),$ $P(X_3|X_1,\omega)\}$. If $\sum_\omega P(\omega)[I_\omega(X_2,X_1)+I_\omega(X_3,X_1)]>\sum_\omega P(\omega)H_\omega$ (X_1), then from (10), we see that the presence of the variable X_1 decreases $\hat{H}(\omega|X)$, thereby decreasing the upper bound on the Bayes error rate. However, the component probability $P(X_4|\omega)$ does not satisfy Condition 2 because l_4 is an empty set. Therefore, the presence of $P(X_4|\omega)$ certainly increases $\hat{H}(\omega|X)$, thereby increasing the upper bound on the Bayes error rate. Consequently, the variable X_4 may be omitted when approximating $P(X|\omega)$.

To conclude, we correct an important equation that relates dependence tree classification error to Bayes error rate and present its implication on the selection of component probabilities for dependence tree approximation.

APPENDIX A

DERIVATION RELATING BAYES ERROR RATE TO DEPENDENCE TREE CLASSIFICATION ERROR

It is known that

$$H(\omega|X)=H(\omega)-I(X,\omega). \tag{4}$$

Using the definition of mutual information [5], $I(X,\omega)$ in (4) can be expanded as

$$\begin{aligned}H(\omega|X)=H(\omega)&-\sum_{x,\omega}P(x,\omega)\log P(x,\omega)\\&+\sum_{x,\omega}P(x,\omega)\log P(x)+\sum_{x,\omega}P(x,\omega)\log P(\omega).\end{aligned} \tag{5}$$

By the definition of entropy, $\sum_{x,\omega}P(x,\omega)\log P(x)=\sum_x P(x)\log$ $P(x)=-H(X)$ and

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

Therefore, (5) can be written as

$$\begin{aligned}H(\omega|X)&=-H(X)-\sum_{x,\omega}P(x,\omega)\log P(x,\omega)\\&=-H(X)-\sum_\omega P(\omega)\sum_x P(x|\omega)\log(P(x|\omega)P(\omega)).\end{aligned} \tag{6}$$

Using dependence tree approximation in (1), $\log(P(x|\omega)P(\omega))$ in (6) is replaced by $\log(\hat{P}(x|\omega)P(\omega))$ so that

$$\begin{aligned}\hat{H}(\omega|X)&=-H(X)+H(\omega)-\sum_\omega P(\omega)\sum_x P(x|\omega)\\&\quad\sum_{i=1}^n\log P(x_i|x_{j(i)},\omega),0\le j(i)<i\\&=-H(X)+H(\omega)\underbrace{-\sum_\omega P(\omega)\sum_x P(x|\omega)\sum_{i=1,\,j(i)\neq 0}^n\log P(x_i|x_{j(i)},\omega)}_{\text{Term I}}\\&\qquad\underbrace{-\sum_\omega P(\omega)\sum_x P(x|\omega)\sum_{i=1,\,j(i)=0}^n\log P(x_i|x_{j(i)},\omega)}_{\text{Term II}}.\end{aligned} \tag{7}$$

Term I (sign included) in (7) contains the component probabilities of the form $P(x_i|x_{j(i)},\omega)$, $j(i)<i$ and $j(i)\neq 0$. Term II (sign included) contains the remaining component probabilities of the form $P(x_i|x_0,\omega)=P(x_i|\omega)$. Term I can be expanded as

$$\begin{aligned}&-\sum_\omega P(\omega)\sum_x P(x|\omega)\sum_{i=1,\,j(i)\neq 0}^n\left(\log\frac{P(x_i,x_{j(i)}|\omega)}{P(x_i|\omega)P(x_{j(i)}|\omega)}\right)\\&-\sum_\omega P(\omega)\sum_x P(x|\omega)\sum_{i=1,\,j(i)\neq 0}^n\log P(x_i|\omega).\end{aligned} \tag{8}$$

Since $P(x_i,x_{j(i)}|\omega)$ and $P(x_i|\omega)$ are components (marginal distributions) of $P(x|\omega)$, we know that

$$\begin{aligned}&\sum_x P(x|\omega)\log\frac{P(x_i,x_{j(i)}|\omega)}{P(x_i|\omega)P(x_{j(i)}|\omega)}=\\&\sum_{x_i,x_{j(i)}}P(x_i,x_{j(i)}|\omega)\log\frac{P(x_i,x_{j(i)}|\omega)}{P(x_i|\omega)P(x_{j(i)}|\omega)}\text{ and}\\&\sum_x P(x|\omega)\log P(x_i|\omega)=\sum_{x_i}P(x_i|\omega)\log P(x_i|\omega).\end{aligned}$$

Therefore, the expansion in (8) can be rewritten as

$$-\sum_\omega P(\omega)\sum_{i=1,\,j(i)\neq 0}^n I_\omega(X_i,X_{j(i)})+\sum_\omega P(\omega)\sum_{i=1,\,j(i)\neq 0}^n H_\omega(X_i). \tag{9}$$

Expansion (9) shows that each component probability of the form $P(x_i|x_{j(i)},\omega)$, $j(i)\neq 0$, adds $\sum_\omega P(\omega)H_\omega(X_i)$ to $\hat{H}(\omega|X)$. Now, consider Term II in (7). Let there be K component probabilities of the form $P(x_i|x_0,\omega)=P(x_i|\omega)$. Then, Term II can be written as

$$\begin{aligned}&-\sum_\omega P(\omega)\sum_x P(x|\omega)\sum_{i=1}^K\log P(x_i|\omega)=\\&-\sum_\omega P(\omega)\sum_{i=1}^K\sum_{x_i}P(x_i|\omega)\log P(x_i|\omega)=\sum_\omega P(\omega)\sum_{i=1}^K H_\omega(X_i),\end{aligned} \tag{10}$$

where $K\ge 1$ and $K\le n$ from the definition of dependence tree approximation in (1). Equation (10) shows that each component probability of the form $P(x_i|x_0,\omega)$ adds $\sum_\omega P(\omega)H_\omega(X_i)$ to $\hat{H}(\omega|X)$. By substituting (9) and (10) for Term I and Term II, respectively, (7) becomes

$$\hat{H}(\omega|X) = H(\omega) - H(X) - \sum_{\omega} P(\omega) \sum_{i=1,\, j(i)\neq 0}^{n} I_{\omega}(X_i, X_{j(i)}) + \sum_{\omega} P(\omega) \sum_{i=1}^{n} H_{\omega}(X_i).$$

ACKNOWLEDGMENTS

The authors thank all of the reviewers for their critical comments, which significantly improved the paper. In particular, the authors thank Reviewer 1 and Reviewer 3 for insightful comments on (1) and the derivation.

REFERENCES

[1] S.K.M. Wong and F.C.S. Poon, "Comments on Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 3, pp. 333-335, Mar. 1989.

[2] C.K. Chow and C.N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Trans. Information Theory*, vol. 14, pp. 462-467, May 1968.

[3] J.B. Kruskal Jr., "On the Shortest Spanning Subtree of a Graph and the Travelling Saleman Problem," *Proc. Conf. Am. Math. Soc.*, vol. 7, pp. 48-50, 1956.

[4] M.E. Hellman and J. Raviv, "Probability of Error, Equivocation, and the Chernoff Bound," *IEEE Trans. Information Theory*, vol. 16, pp. 368-372, May 1970.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. Wiley Interscience, 1991.

[6] H. Avi-ltzhak and T. Diep, "Arbitrarily Tight Upper and Lower Bounds on the Bayesian Probability of Error," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89-91, Jan. 1996.