On Guo and Nixon's Criterion for Feature Subset Selection: Assumptions, Implications, and Alternative Options

Article *in* IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans · June 2010

DOI: 10.1109/TSMCA.2009.2036935 · Source: IEEE Xplore

CITATIONS
4

READS
30

4 authors:

Kiran Balagani
New York Institute of Technology
43 PUBLICATIONS **1,287** CITATIONS
SEE PROFILE

Vir V. Phoha
Louisiana Tech University
216 PUBLICATIONS **3,151** CITATIONS
SEE PROFILE

Sundararaj Iyengar
Florida International University
685 PUBLICATIONS **11,602** CITATIONS
SEE PROFILE

Narayanaswamy Balakrishnan
Indian Institute of Science
209 PUBLICATIONS **2,351** CITATIONS
SEE PROFILE

On Guo and Nixon's Criterion for Feature Subset Selection: Assumptions, Implications, and Alternative Options

Kiran S. Balagani, *Member, IEEE*,
Vir V. Phoha, *Senior Member, IEEE*,
S. S. Iyengar, *Fellow, IEEE*, and N. Balakrishnan

***Abstract*—Guo and Nixon proposed a feature selection method based on maximizing $I(\mathbf{x};Y)$, the multidimensional mutual information between feature vector $\mathbf{x}$ and class variable Y. Because computing $I(\mathbf{x};Y)$ can be difficult in practice, Guo and Nixon proposed an approximation of $I(\mathbf{x};Y)$ as the criterion for feature selection. We show that Guo and Nixon's criterion originates from approximating the joint probability distributions in $I(\mathbf{x};Y)$ by second-order product distributions. We remark on the limitations of the approximation and discuss computationally attractive alternatives to compute $I(\mathbf{x};Y)$.**

***Index Terms*—Entropic spanning graphs, feature selection, mutual information, Parzen window.**

Notation and Formulas

Here, we briefly give the notation and formulas used in our paper. For readers' convenience, we retain the notation of [1] as much as possible.

Let $\mathbf{x}$ denote a d-dimensional feature vector. Let X_i and X_j denote the ith and jth features in $\mathbf{x}$, respectively. Let Y denote the class variable representing classes $\{y_1,\ldots,y_k\}$. We refer to the probability distributions having two features as second-order probability distributions (e.g., $P(X_i,X_j)$ and $P(X_i,X_j|Y)$). We refer to the probability distributions with more than two features as higher order probability distributions.

The mutual information [2] between feature X_i and class variable Y is

$$I(X_i;Y)=\sum_{x_i\in X_i}\sum_{y\in Y}P(x_i,y)\log\frac{P(x_i,y)}{P(x_i)P(y)} \tag{1}$$

and the mutual information between features X_i and X_j is

$$I(X_i;X_j)=\sum_{x_i\in X_i}\sum_{x_j\in X_j}P(x_i,x_j)\log\frac{P(x_i,x_j)}{P(x_i)P(x_j)}. \tag{2}$$

The conditional mutual information [2] between features X_i and X_j, given class variable Y, is

$$I(X_i;X_j|Y)=\sum_{X_i\in x_i}\sum_{X_j\in x_j}\sum_{y\in Y}P(x_i,x_j,y)\times\log\frac{P(x_i,x_j|y)}{P(x_i|y)P(x_j|y)}. \tag{3}$$

The multidimensional mutual information between feature vector $\mathbf{x}$ and class variable Y is

$$\begin{aligned}I(\mathbf{x};Y)&=I(X_1,\ldots,X_d;Y)\\&=\sum_{x_1,\ldots,x_d}\sum_{y\in Y}P(x_1,\ldots,x_d,y)\\&\quad\times\log\frac{P(x_1,\ldots,x_d,y)}{P(x_1,\ldots,x_d)P(y)}.\end{aligned} \tag{4}$$

I. Introduction

Guo and Nixon [1] demonstrated the effectiveness of a sequential feature subset selection method for human gait recognition application. The feature selection method involves selecting a feature subset that maximizes $I(\mathbf{x};Y)$. Using $I(\mathbf{x};Y)$ as the feature selection criterion is backed by two reasons: 1) Because $I(\mathbf{x};Y)$ is a measure of the reduction of uncertainty in class Y due to the knowledge of feature vector $\mathbf{x}$, selecting features that maximize $I(\mathbf{x};Y)$, from an information-theoretic perspective, translates into selecting those features that contain the maximum information about class Y, and 2) maximizing $I(\mathbf{x};Y)$ minimizes a lower bound on the Bayes classification error. Guo and Nixon proved 2) by expanding class conditional entropy $H(Y|X)$ in Fano's inequality (see [1, eq. (14)]).

In practice, finding a feature subset that maximizes $I(\mathbf{x};Y)$ has two problems: 1) It requires an exhaustive "combinatorial" search over the feature space, and 2) it requires estimation of higher order joint probability distributions—the number of joint probability distributions required to estimate $I(\mathbf{x};Y)$ grows exponentially with the number of features (see [3] and [4] for discussions on the complexity issues). Guo and Nixon proposed a second-order approximation to $I(\mathbf{x};Y)$ as the feature selection criterion. The approximation is given as

$$\begin{aligned}I(\mathbf{x};Y)&\approx\hat{I}(\mathbf{x};Y)\\&=\sum_i I(X_i;Y)-\sum_i\sum_{j>i}I(X_i;X_j)\\&\quad+\sum_i\sum_{j>i}I(X_i;X_j|Y).\end{aligned} \tag{5}$$

By using $\hat{I}(\mathbf{x};Y)$ instead of $I(\mathbf{x};Y)$, Guo and Nixon were able to find a subset of informative features by implementing a greedy "pick-one-feature-at-a-time" selection procedure. Given n features, out of which m are to be selected ($m<n$), Guo and Nixon's procedure proceeds in two steps: 1) Select the first feature $X'_{\max}$ that maximizes $I(X';Y)$, and 2) select $m-1$ subsequent features that maximize the criterion in (5), i.e., select the second feature $X''_{\max}$ that maximizes $I(X'';Y)-I(X'';X'_{\max})+I(X'';X'_{\max}|Y)$, select the third feature $X'''_{\max}$ that maximizes $I(X''';Y)-I(X''';X'_{\max})-I(X''';X''_{\max})+I(X''';X'_{\max}|Y)+I(X''';X''_{\max}|Y)$, and so on. The first step of the procedure computes mutual information n times. The second step computes mutual information and conditional mutual information $2\sum_{k=1}^{m-1}k(n-k)$ times, which is on the order of $O(m^3+nm^2+m)$. In each step, computing mutual information

Manuscript received February 11, 2009; revised June 13, 2009. First published January 22, 2010; current version published April 14, 2010. This work was supported in part by the Louisiana Board of Regents under PKSFI Grant LEQSF(2007-12)-ENH-PKSFI-PRS-03. This paper was recommended by Associate Editor J. Wu.

K. S. Balagani and V. V. Phoha are with Louisiana Tech University, Ruston, LA 71270 USA (e-mail: ksb011@latech.edu; phoha@latech.edu).

S. S. Iyengar is with Louisiana State University, Baton Rouge, LA 70809 USA (e-mail: iyengar@bit.csc.lsu.edu).

N. Balakrishnan is with the Indian Institute of Science, Bangalore 560012, India (e-mail: balki@aero.iisc.ernet.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2009.2036935

and conditional mutual information requires estimation of r^2 and r^3 probability distributions, respectively, assuming that the features and the class variable take on "r" discrete values.

II. Purpose and Contributions of Our Paper

A. Purpose

The purpose of our paper is to complement [1] by doing the following: 1) presenting the limitations of Guo and Nixon's feature selection criterion; 2) revealing the root cause of the limitations; and 3) pointing to alternatives that mitigate the limitations. Because Guo and Nixon's feature selection criterion is equally applicable for many classification problems beyond gait recognition, we opine that our remarks, along with the merits demonstrated in [1], will aid interested practitioners in weighing the pros and cons of using the criterion.

B. Contributions

We present two remarks on $\hat{I}(\mathbf{x};Y)$. A brief description of the remarks follow.

1) Approximating $I(\mathbf{x};Y)$ as $\hat{I}(\mathbf{x};Y)$ [see (5)] means that joint probability distributions $P(\mathbf{x},Y)$ and $P(\mathbf{x})$ are approximated as products of second-order probability distributions. We prove the exact forms of the second-order product distributions that lead to approximation $\hat{I}(\mathbf{x};Y)$. In the first remark, we discuss two limitations of using $\hat{I}(\mathbf{x};Y)$ for feature selection.
2) The primary reason for using approximation $\hat{I}(\mathbf{x};Y)$ for feature selection instead of directly using multidimensional mutual information $I(\mathbf{x};Y)$ is that $I(\mathbf{x};Y)$ requires estimation of the joint probability distribution of features—the number of joint distributions to be estimated rises exponentially with the number of features. While this argument is true when a histogram-based estimate of $I(\mathbf{x};Y)$ is used, there are alternate ways to estimate $I(\mathbf{x};Y)$ that do not incur the exponential complexity burden. In the second remark, we briefly discuss two such estimators of $I(\mathbf{x};Y)$ and point to relevant studies in the literature.

III. Proof of How $\hat{I}(\mathbf{x};Y)$ Approximates $I(\mathbf{x};Y)$ and Remarks

In this section, we present a proposition and two remarks. Proposition 1 shows how $\hat{I}(\mathbf{x};Y)$ approximates $I(\mathbf{x};Y)$ by using second-order product approximations of joint probability distributions $P(\mathbf{x},Y)$ and $P(\mathbf{x})$. Furthermore, Proposition 1 gives the root cause of the limitations raised in Remark 1.

Proposition 1: Let $\mathbf{x}=(X_1,\ldots,X_m)$ denote an m-dimensional feature vector. Let $P(\mathbf{x})$ be the joint probability distribution of $\mathbf{x}$, and let $P(\mathbf{x},Y)$ be the joint probability distribution of the features and the class variable. Let $\hat{P}(\mathbf{x})$ be a second-order product approximation of $P(\mathbf{x})$. Let $\hat{P}(\mathbf{x},Y)$ be a second-order product approximation of $P(\mathbf{x},Y)$. Multidimensional mutual information $I(\mathbf{x};Y)$ becomes $\hat{I}(\mathbf{x};Y)$ when

$$\hat{P}(\mathbf{x}) = P(X_1)P(X_2|X_1)\frac{P(X_3|X_2)P(X_3|X_1)}{P(X_3)}\cdots \frac{P(X_m|X_1)P(X_m|X_2)\cdots P(X_m|X_{m-1})}{[P(X_m)]^{m-2}}$$

$$\hat{P}(\mathbf{x},Y) = P(Y)P(X_1|Y)P(X_2|X_1,Y) \times \frac{P(X_3|X_1,Y)P(X_3|X_2,Y)}{P(X_3|Y)}\cdots \frac{P(X_m|X_1,Y)\cdots P(X_m|X_{m-1},Y)}{[P(X_m|Y)]^{m-2}}.$$

Proof: By the multiplication rule of probability, we expand $P(\mathbf{x},Y)=P(X_1,\ldots,X_m,Y)$ as

$$P(Y)P(X_1|Y)P(X_2|X_1,Y)P(X_3|X_1,X_2,Y) \times P(X_4|X_1,X_2,X_3,Y)\cdots P(X_m|X_1,\ldots,X_{m-1},Y). \quad (6)$$

If we assume that the conditioning variables X_1 and X_2 in $P(X_3|X_1,X_2,Y)$ are independent, we have (7), shown at the bottom of the page. In a similar fashion, if we assume that the conditioning variables $X_1,\ldots,X_{m-1}$ are independent, then $P(X_m|X_1,\ldots,X_{m-1},Y)$ becomes (8), shown at the bottom of the next page. By assuming independence among conditioning variables, each of the higher order probability terms in (6), i.e., $P(X_3|X_1,X_2,Y)$ through $P(X_m|X_1,\ldots,X_{m-1},Y)$, can be reduced to products of second-order distributions [as done in (7) and (8)], so that

$$\begin{aligned} P(\mathbf{x},Y) &\approx \hat{P}(\mathbf{x},Y) \\ &= P(Y)P(X_1|Y)P(X_2|X_1,Y) \times \frac{P(X_3|X_1,Y)P(X_3|X_2,Y)}{P(X_3|Y)} \\ &\quad \times \frac{P(X_4|X_1,Y)P(X_4|X_2,Y)P(X_4|X_3,Y)}{[P(X_4|Y)]^2}\cdots \frac{P(X_m|X_1,Y)\cdots P(X_m|X_{m-1},Y)}{[P(X_m|Y)]^{m-2}}. \end{aligned} \quad (9)$$

Consider joint distribution $P(\mathbf{x})=P(X_1,\ldots,X_m)$. By the multiplication rule

$$P(X_1,\ldots,X_m) = P(X_1)P(X_2|X_1)P(X_3|X_2,X_1) \times P(X_4|X_3,X_2,X_1)\cdots P(X_m|X_1,X_2,\ldots,X_{m-1}). \quad (10)$$

$$\begin{aligned} P(X_3|X_1,X_2,Y) &= \frac{P(X_1,X_2,X_3,Y)}{P(X_1,X_2,Y)} = \frac{P(X_3|Y)P(X_1|X_3,Y)P(X_2|X_3,X_1,Y)}{P(X_1|Y)P(X_2|X_1,Y)} \\ &= \frac{P(X_3|Y)P(X_1|X_3,Y)P(X_2|X_3,Y)}{P(X_1|Y)P(X_2|Y)} = \frac{P(X_3|Y)P(X_1,X_3,Y)P(X_2,X_3,Y)}{P(X_1|Y)P(X_2|Y)P(X_3,Y)P(X_3,Y)} \\ &= \frac{P(X_3|Y)P(Y)P(X_1|Y)P(X_3|X_1,Y)P(Y)P(X_2|Y)P(X_3|X_2,Y)}{P(X_1|Y)P(X_2|Y)P(Y)P(X_3|Y)P(Y)P(X_3|Y)} \\ &= \frac{P(X_3|X_1,Y)P(X_3|X_2,Y)}{P(X_3|Y)} \end{aligned} \quad (7)$$

Again, by assuming independence among conditioning variables, the higher order probability terms in (10), i.e., $P(X_3|X_2, X_1)$ through $P(X_m|X_1, X_2, \ldots, X_{m-1})$, can be reduced to products of second-order distributions, so that

$$\begin{aligned} P(\mathbf{x}) &\approx \hat{P}(\mathbf{x}) \\ &= P(X_1)P(X_2|X_1)\frac{P(X_3|X_2)P(X_3|X_1)}{P(X_3)} \\ &\quad \times \frac{P(X_4|X_3)P(X_4|X_2)P(X_4|X_1)}{[P(X_4)]^2} \cdots \\ &\quad \frac{P(X_m|X_1)P(X_m|X_2)\cdots P(X_m|X_{m-1})}{[P(X_m)]^{m-2}}. \end{aligned} \tag{11}$$

Let X_i and X_j $(i \neq j)$ be any two features in feature vector $\mathbf{x} = (X_1, \ldots, X_m)$. Then, by marginalization

$$\begin{aligned} &\sum_{\mathbf{x}}\sum_{Y} P(\mathbf{x}, Y)\log\left(\frac{P(X_i|X_j, Y)}{P(X_i|X_j)}\right) \\ &= \sum_{X_1,\ldots,X_m}\sum_{Y} P(X_1, \ldots, X_m, Y) \\ &\quad \times \log\left(\frac{P(X_i, X_j|Y)}{P(X_i|Y)P(X_j|Y)}\frac{P(X_i)P(X_j)}{P(X_i, X_j)}\frac{P(X_i|Y)}{P(X_i)}\right) \\ &= \sum_{X_i}\sum_{Y} P(X_i, Y)\log\left(\frac{P(X_i|Y)}{P(X_i)}\right) \\ &\quad + \sum_{X_i,X_j}\sum_{Y} P(X_i, X_j, Y)\log\left(\frac{P(X_i, X_j|Y)}{P(X_i|Y)P(X_j|Y)}\right) \\ &\quad - \sum_{X_i,X_j} P(X_i, X_j)\log\left(\frac{P(X_i, X_j)}{P(X_i)P(X_j)}\right) \\ &= I(X_i; Y) + I(X_i; X_j|Y) - I(X_i; X_j). \end{aligned} \tag{12}$$

By substituting $P(\mathbf{x}, Y)$ with $\hat{P}(\mathbf{x}, Y)$ [from (9)] and $P(\mathbf{x})$ with $\hat{P}(\mathbf{x})$ [from (11)] and by using the result of (12), we have (13), shown at the bottom of the page. Thus, we prove Proposition 1 by showing that $I(\mathbf{x}; Y)$ becomes Guo and Nixon's criterion $\hat{I}(\mathbf{x}; Y)$ when the joint probability distributions $P(\mathbf{x})$ and $P(\mathbf{x}, Y)$ in $I(\mathbf{x}; Y)$ are approximated by $\hat{P}(\mathbf{x})$ and $\hat{P}(\mathbf{x}, Y)$, respectively.

Remark 1: There are two limitations of using approximation $\hat{I}(\mathbf{x}, Y)$ [see (5)] as a feature selection criterion.

The first limitation, attributed to the higher order independence assumptions made in (11), is that *criterion $\hat{I}(\mathbf{x}, Y)$ does not check for third and higher order dependencies between the features*. With a hypothetical example, we explain the negative effect of this limitation. Let X_1, X_2, X_3, and X_4 be four features, out of which we wish to select three. We assume a scenario in which X_1 and X_2 are the features already selected by the criterion; X_3 and X_4 have equal associations with class variable Y, i.e., $I(X_3; Y) = I(X_4; Y)$, $I(X_1; X_3|Y) = I(X_1; X_4|Y)$, and $I(X_2; X_3|Y) = I(X_2; X_4|Y)$; and $I(X_1, X_2; X_3) \gg I(X_1; X_3) + I(X_2; X_3)$, meaning that the dependence that X_3 has with (X_1, X_2) jointly considered is considerably greater than the sum of the dependencies that X_3 has individually with X_1 and X_2. Now, to select the third feature, the criterion chooses between X_3 and X_4 by comparing $\phi_{13} = I(X_1; X_3) + I(X_2; X_3)$ and $\phi_{14} = I(X_1; X_4) + I(X_2; X_4)$, while completely ignoring the possibility that $I(X_1, X_2; X_3)$ could be significantly greater than ϕ_{13} and ϕ_{14}, in which case X_4 may be a better choice.

The second limitation, attributed to the higher order independence assumption made in (9), is that *criterion $\hat{I}(\mathbf{x}, Y)$ does not consider third and higher order associations between the features and the class*. We explain this limitation with a hypothetical example. Our example is a 3-D extension of the XOR classification problem [5]. Let X_1, X_2, and X_3 denote three binary-valued (0/1) features, and let $Y = \{y_1, y_2\}$ denote the class labels associated with the data points in (X_1, X_2, X_3) 3-D feature space. Fig. 1 shows eight data points lying on the corners of a cube in (X_1, X_2, X_3) feature space such that no two data points of the same class share an edge. From Fig. 1, for all values of X_1, X_2, X_3, and Y, we have $P(X_1) = P(X_2) = P(X_3) = P(Y) = 1/2$, $P(X_1, Y) = P(X_2, Y) = P(X_3, Y) = 1/4$,

---

$$\begin{aligned} P(X_m|X_1, \ldots, X_{m-1}, Y) &= \frac{P(X_1, \ldots, X_m, Y)}{P(X_1, \ldots, X_{m-1}, Y)} = \frac{P(X_m|Y)P(X_1|X_m, Y)\cdots P(X_{m-1}|X_1, X_2, \ldots, X_{m-2}, X_m, Y)}{P(X_1|Y)\cdots P(X_{m-1}|X_1, X_2, \ldots, X_{m-2}, Y)} \\ &= \frac{P(X_m|Y)P(X_1|X_m, Y)\cdots P(X_{m-1}|X_m, Y)}{P(X_1|Y)\cdots P(X_{m-1}|Y)} \\ &= \frac{P(X_m|Y)P(Y)P(X_1|Y)P(X_m|X_1, Y)\cdots P(Y)P(X_{m-1}|Y)P(X_m|X_{m-1}, Y)}{P(X_1|Y)\cdots P(X_{m-1}|Y)\,[P(X_m, Y)]^{m-1}} \\ &= \frac{P(X_m|X_1, Y)\cdots P(X_m|X_{m-1}, Y)}{[P(X_m|Y)]^{m-2}} \end{aligned} \tag{8}$$

$$\begin{aligned} I(\mathbf{x}, Y) &= \sum_{\mathbf{x}}\sum_{Y} P(\mathbf{x}, Y)\log\frac{P(\mathbf{x}, Y)}{P(\mathbf{x})P(Y)} \approx \sum_{\mathbf{x}}\sum_{Y} P(\mathbf{x}, Y)\log\frac{\hat{P}(\mathbf{x}; Y)}{\hat{P}(\mathbf{x})P(Y)} \\ &= \sum_{\mathbf{x}}\sum_{Y} P(\mathbf{x}, Y)\log\left(\frac{P(X_1, Y)}{P(X_1)P(Y)}\frac{P(X_2|X_1, Y)}{P(X_2|X_1)}\cdots\frac{P(X_m|X_1, Y)\cdots P(X_m|X_{m-1}, Y)\,[P(X_m)]^{m-2}}{P(X_m|X_1)\cdots P(X_m|X_{m-1})\,[P(X_m|Y)]^{m-2}}\right) \\ &= \hat{I}(\mathbf{x}, Y) \end{aligned} \tag{13}$$

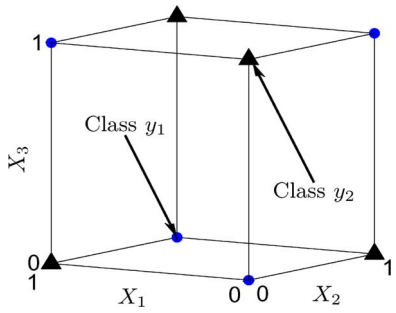

Fig. 1. Three-dimensional data points lying on the corners of a cube such that no two points of the same class share an edge. In this arrangement, data points of classes ("•") y_1 and ("Δ") y_2 become separable only in the $\mathbf{x} = (X_1, X_2, X_3)$ joint feature space.

$P(X_1, X_2) = P(X_1, X_3) = P(X_2, X_3) = 1/4$, $P(X_1, X_2, Y) = P(X_1, X_3, Y) = P(X_2, X_3, Y) = P(X_1, X_2, X_3) = 1/8$, and $P(X_1, X_2, X_3, Y) = 1/8$ or 0. Under this arrangement, Guo and Nixon's feature selection criterion does not select any of the features X_1, X_2, and X_3 because all second-order terms used for computing $\hat{I}(\mathbf{x}; Y)$, i.e., $I(X_1; Y)$, $I(X_2; Y)$, $I(X_3; Y)$, $I(X_1; X_2|Y)$, $I(X_1; X_3|Y)$, and $I(X_2; X_3|Y)$, equal zero assuming base 2 logarithm. In contrast, multidimensional mutual information $I(X_1, X_2, X_3; Y) = 1$, being considerably greater than the second-order terms of Guo and Nixon's criterion, indicates that higher class separability can be achieved by considering all the three features X_1, X_2, and X_3 together and therefore favors their selection (a similar argument motivated Kwak and Choi's work in [6]).

We opine that pointing out the second limitation is important also for the following reason. In the literature, mutual-information-based feature selection methods like the one presented by Guo and Nixon [1] are generally presented as "filter" approaches to feature selection, implying that Guo and Nixon's criterion can comply with any classifier. Although this is partly true, consider using Guo and Nixon's criterion when Bayesian networks [7] are used for subsequent classification. These networks rely on (and are sensitive to) the dependencies between the features and the class. However, because Guo and Nixon's criterion does not check for higher order dependencies between the features and the class [see (9)], the criterion risks the elimination of features that might indeed be useful for classification using a Bayesian network. In such situations, a practitioner will benefit by knowing the underlying assumptions of Guo and Nixon's criterion.

The root cause of the limitations identified in Remark 1 is the higher order independence assumption made on the joint probability distribution of features [see (9) and (11)]. By directly estimating $I(\mathbf{x}; Y)$, we can avoid making the assumption and thus circumvent the limitations. In the next remark, we briefly discuss two computationally practical ways to compute $I(\mathbf{x}; Y)$ for feature selection.

Remark 2: Multidimensional mutual information $I(\mathbf{x}; Y)$ can be estimated using a Parzen-window-based "plug-in" estimate or an entropic-spanning-graph-based "bypass" estimate. These estimates circumvent the exponential complexity incurred when $I(\mathbf{x}; Y)$ is estimated using histograms.

Because $I(\mathbf{x}; Y) = H(Y) - H(Y|\mathbf{x})$, estimates of entropy [2] can as well be used to estimate $I(\mathbf{x}; Y)$. There are two types of entropy estimates: 1) "plug-in" estimates [8] and 2) "bypass" [12] estimates. An example of a plug-in estimate is the integral estimate [8], which estimates the joint probability density in $H(Y|\mathbf{x})$ with a kernel density estimator (note that a histogram-based estimate of $I(\mathbf{x}; Y)$ is also a type of plug-in estimate). Kwak and Choi [6] demonstrated a sequential forward feature selection procedure based on maximizing $I(\mathbf{x}; Y)$, which is calculated by estimating entropies using Parzen windows. To select m out of n available features, Kwak and Choi's procedure proceeds as follows: Iteratively select m features such that each selected feature maximizes the multidimensional mutual information between the features and the class, i.e., the first feature $X'_{\max}$ maximizes $I(X'; Y)$, the second feature $X''_{\max}$ maximizes $I(X'', X'_{\max}; Y)$, the third feature $X'''_{\max}$ maximizes $I(X''', X''_{\max}, X'_{\max}; Y)$, and so on. In each iteration, the dimensionality of the Parzen window increases by one feature. Kwak and Choi's procedure estimates multidimensional mutual information $(nm - (m^2/2) + (m/2))$ times. The complexity of estimating multidimensional mutual information using a Parzen window is $O(N^2 m)$, where N is the number of training samples.

In practice, estimating entropies (and mutual information) with Parzen windows requires careful selection of smoothing parameters and window functions [9]. Because the underlying class conditional densities of features are largely unknown in a nonparametric feature selection setting, enforcing incautious assumptions on the smoothing parameter or the functional form of the window can lead to severe over- or underparameterization of the densities. Furthermore, the demand for the number of training samples increases exponentially with the dimensionality of the Parzen window [10], which may be unacceptable for the problem at hand. These reasons motivate the direct estimation of entropy through a "bypass" estimator.

A bypass entropy estimate is obtained by constructing an entropic graph [11], which is a minimal graph spanning the training samples. Examples of entropic graphs include minimal spanning trees, Steiner trees, and minimal matching graphs (see [12]). An entropic graph, however, does not directly estimate Shannon's entropy $H(.)$ but estimates Renyi's α-entropy $H_\alpha(.)$ [2], of which Shannon's entropy is a special case when $\alpha = 1$. Let $\mathbf{\Gamma} = \{\mathbf{X}_1, \ldots, \mathbf{X_N}\}$ be a set of N m-dimensional i.i.d. training samples. The length $L(\mathbf{\Gamma})$ of the entropic graph on the samples in $\mathbf{\Gamma}$ is given as $L(\mathbf{\Gamma}) = \min\{\sum_e |e|^\gamma\}$, where e denotes an edge in the spanning graph, $|.|$ denotes the Euclidean distance between pairs of samples, and γ is a power (weight) term. The α-entropy is estimated as

$$\hat{H}_\alpha = \frac{1}{1-\alpha}\left[\ln \frac{L(\Gamma)}{n^\alpha} - \ln \beta(\gamma, m)\right] \tag{14}$$

where $\alpha = (m - \gamma)/m$ and $\beta(\gamma, m)$ is equal to $(\gamma/2)\ln(m/2\pi e)$ if a minimal spanning tree is used [11], [12]. Note that $\hat{H}_\alpha$ is an asymptotically unbiased and almost surely consistent estimator under certain conditions (see [12]). Bonev *et al.* [13] demonstrated the success of entropic graphs (minimal spanning trees specifically) to estimate $I(\mathbf{x}; Y)$ for feature selection. Like Kwok and Choi [6], Bonev *et al.* use a sequential forward selection procedure to select features. To estimate Shannon's entropy, Bonev *et al.* use (14) to estimate the α-entropy for α values near one and then estimate $\hat{H}(.)$ at $\alpha = 1$ by extrapolation. The computational complexity of estimating $I(\mathbf{x}; Y)$ is $O(m \times n \log n)$, where $n \log n$ is the complexity of constructing a minimal spanning tree.

IV. Conclusion

We have presented two remarks on Guo and Nixon's mutual-information-based criterion for feature subset selection. In the first remark, we explained the limitations of Guo and Nixon's criterion and showed (using Proposition 1) that the limitations arise from the higher order independence assumption on the joint probability distribution of features. In the second remark, we discussed two ways to directly estimate the multidimensional mutual information between the features and the class variable for feature subset selection. Our remarks intend to complement the merits of Guo and Nixon's criterion discussed in their paper [1].

ACKNOWLEDGMENT

The authors thank the six anonymous reviewers for their constructive suggestions. The authors also thank C. Duncan of Computer Science, Louisiana Tech University, for the benefits of discussions pertaining to Fig. 1.

REFERENCES

[1] B. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 1, pp. 36–46, Jan. 2009.

[2] T. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 1999, ser. Wiley Series in Telecommunications.

[3] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[4] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[6] N. Kwak and C. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[7] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2/3, pp. 131–163, Nov. 1997.

[8] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[9] R. P. W. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Trans. Comput.*, vol. C-25, no. 11, pp. 1175–1179, Nov. 1976.

[10] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.

[11] A. O. Hero, III, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Process. Mag.*, vol. 19, no. 5, pp. 85–95, Sep. 2002.

[12] A. O. Hero, III and O. J. J. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1921–1938, Sep. 1999.

[13] B. Bonev, F. Escolano, and M. Cazorla, "Feature selection, mutual information, and the classification of high-dimensional patterns," *Pattern Anal. Appl.*, vol. 11, no. 3/4, pp. 309–319, Sep. 2008.