

On the Feature Selection Criterion Proposed in ‘Gait Feature Subset Selection by Mutual Information’

Kiran S. Balagani¹ Vir. V. Phoha¹ S. S. Iyengar² N. Balakrishnan³

¹Louisiana Tech University
Ruston, LA 71270, USA
{ksb011,phoha}@latech.edu

²Louisiana State University
Baton Rouge, LA 70809, USA
iyengar@bit.csc.lsu.edu

³Indian Institute of Science
Bangalore, 560012, India
balki@aero.iisc.ernet.in

Abstract

Recently, Guo and Nixon [1] 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 alternatives to compute $I(\mathbf{x}; Y)$ without sacrificing computational economy.

Index Terms–Feature selection, mutual information, Parzen window, entropic spanning graphs.

1 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 denote the i^{th} feature in \mathbf{x} . Let Y denote the class variable representing the classes $\{y_1, \dots, y_k\}$. 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 the class variable Y is

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

the mutual information between the features X_i and X_j is

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

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

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

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

$$I(\mathbf{x}; Y) = I(X_1, \dots, X_d; Y) = \sum_{x_1, \dots, x_d} \sum_{y \in Y} P(x_1, \dots, x_d, y) \log \frac{P(x_1, \dots, x_d, y)}{P(x_1, \dots, x_d)P(y)}. \quad (4)$$

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

In practice, finding a subset of features that maximizes $I(\mathbf{x}; Y)$ has two problems: (a) it requires an exhaustive “combinatorial” search over the feature space, and (b) it requires estimating higher order joint probability distributions—computational complexity of estimating joint probability distributions grows exponentially with the number of features. (See studies [3, 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

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

By using $\hat{I}(\mathbf{x}; Y)$ instead of $I(\mathbf{x}; Y)$, Guo and Nixon were able to restrict computations to second-order joint probability distributions and were able to find a subset of informative features by implementing a greedy ‘pick-one-feature-at-a-time’ selection strategy.

3 Purpose and Contributions of our Paper

3.1 Purpose

The purpose of our paper is to complement [1] by (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.

3.2 Contributions

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

- Approximating $I(\mathbf{x}; Y)$ as $\hat{I}(\mathbf{x}; Y)$ (see (5)) means that the 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 the approximation $\hat{I}(\mathbf{x}; Y)$. In the first remark, we discuss two limitations of using $\hat{I}(\mathbf{x}; Y)$ for feature selection.
- The primary reason for using the approximation $\hat{I}(\mathbf{x}; Y)$ for feature selection instead of directly using the multidimensional mutual information $I(\mathbf{x}; Y)$ is that, $I(\mathbf{x}; Y)$ requires calculating the joint probability distribution of features, whose computational complexity 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.

4 Remarks

In this section we present the remarks. Proposition 1 gives the second-order probability distributions used to approximate $I(\mathbf{x}; Y)$ as $\hat{I}(\mathbf{x}; Y)$. Proposition 1 supports the limitations raised in Remark 1.

Proposition 1. *Let $\mathbf{x} = (X_1, \dots, X_m)$ denote an m -dimensional feature vector. Let $P(\mathbf{x})$ be the joint probability distribution of \mathbf{x} and $P(\mathbf{x}, Y)$ 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 $P(\mathbf{x}, Y)$. The multidimensional mutual information $I(\mathbf{x}; Y)$ becomes $\hat{I}(\mathbf{x}; Y)$ when*

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

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

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

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

If we assume that the conditioning variables X_1 and X_2 in $P(X_3|X_1, X_2, Y)$ are independent, then

$$P(X_3|X_1, X_2, Y) = \frac{P(X_3|Y)P(X_1|X_3, Y)P(X_2|X_3, Y)}{P(X_1|Y)P(X_2|Y)} = \frac{P(X_3|X_1, Y)P(X_3|X_2, Y)}{P(X_3|Y)}. \quad (7)$$

Similarly, if we assume that the conditioning variables X_1, \dots, X_{m-1} in $P(X_m|X_1, \dots, X_{m-1}, Y)$ are independent, then

$$P(X_m|X_1, \dots, X_{m-1}, Y) = \frac{P(X_m|X_1, Y) \cdots P(X_m|X_{m-1}, Y)}{[P(X_m|Y)]^{m-2}}. \quad (8)$$

By assuming independence among conditioning variables, the higher order probability terms in (6), i.e., $P(X_3|X_1, X_2, Y)$ through $P(X_m|X_1, \dots, X_{m-1}, Y)$, can be reduced to products of second-order distributions (as done in (7) and (8)), so that

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

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

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

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

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

Let X_i and X_j be any two components of the feature vector \mathbf{x} . Then,

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

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 using the result of (12), we get

$$I(\mathbf{x}, Y) = \sum_{\mathbf{x}} \sum_Y P(\mathbf{x}, Y) \log \frac{P(\mathbf{x}, Y)}{P(\mathbf{x})P(Y)} \approx \sum_{\mathbf{x}} \sum_Y P(\mathbf{x}, Y) \log \frac{\hat{P}(\mathbf{x}, Y)}{\hat{P}(\mathbf{x})P(Y)} = \hat{I}(\mathbf{x}, Y).$$

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 the 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 *the criterion $\hat{I}(\mathbf{x}, Y)$ does not check for third and higher order dependencies between 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 the 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 dependency that X_3 has with (X_1, X_2) jointly considered is considerably greater than the sum of the dependencies 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.

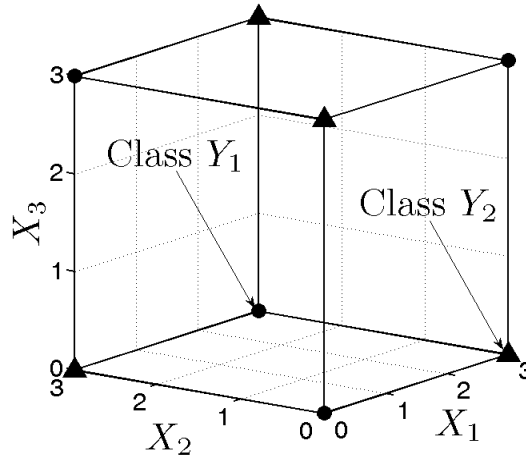


Figure 1: Three-dimensional data points lying on the corners of a cube such that no two points of the same class share an edge. In this arrangement, data points of classes Y_1 (“•”) and Y_2 (“Δ”) become separable only in the $\mathbf{x} = (X_1, X_2, X_3)$ joint feature space.

The second limitation, attributed to the higher order independence assumption made in (9), is that *the 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 three-dimensional extension of the well-known XOR classification problem [5]. Let X_1 , X_2 , and X_3 three features. Let Y_1 and Y_2 denote two classes. Figure 1 shows eight three-dimensional data points lying on the corners of a cube in \mathfrak{R}^3 , such that no two data points of the same class share an edge. In this arrangement, all three features are required to achieve class separability, which from an information-theoretic perspective, translates to $I(X_1, X_2, X_3; Y)$ being considerably greater than the lower-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 $I(X_2; X_3|Y)$, all of which equal zero assuming that each feature is partitioned into four bins $\{0, 1, 2, 3\}$. (A similar argument motivated Kwak and Choi’s work in [6].)

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: The multidimensional mutual information $I(\mathbf{x}; Y)$ can be estimated using a Parzen window based “plug-in” estimate or an entropic spanning graph based “by-pass” estimate. These estimates circumvent the exponential computational 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 [7] and 2) “by-pass” [8] estimates. An example of a plug-in estimate is the integral estimate [7], which replaces 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 feature selection method based on maximizing $I(\mathbf{x}; Y)$, which was estimated by a Parzen window based integral type estimate. The computational complexity of estimating $I(\mathbf{x}; Y)$ is $O(n^2m)$, where n is the number of training feature vectors and m is the dimensionality.

On the other hand, a by-pass entropy estimate is obtained by constructing an entropic spanning graph [8]. An entropic spanning graph, however, does not directly estimate Shannon’s entropy $H(\cdot)$, but estimates Renyi’s α -entropy $H_\alpha(\cdot)$ [2], of which Shannon’s entropy is a special case when $\alpha = 1$. Let $X_n = \{X_1, \dots, X_n\}$ be a set of n m -dimensional i.i.d. training vectors. Let $L(X_n) = \sum_e |e|^\gamma$, where e denotes an edge in the minimal spanning tree constructed on the vectors in X_n , $|\cdot|$ denotes the Euclidean norm, and γ is a power term. The α -entropy is estimated as

$$\hat{H}_\alpha = \frac{1}{1 - \alpha} \left[\ln \frac{L(X_n)}{n^\alpha} - \ln \beta(\gamma, m) \right], \quad (13)$$

where $\alpha = (m - \gamma)/m$ and $\beta(\gamma, m)$ is equal to $\frac{\gamma}{2} \ln(m/2\pi e)$ if a minimal spanning tree construction is used. Shannon’s entropy is estimated by first estimating H_α for α values near 1 and then extrapolation is used to find $H(\cdot)$ at $\alpha = 1$. Bonev et al. [9] demonstrated the

success of entropic spanning graphs to estimate $I(\mathbf{x}; Y)$ for feature selection. 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.

5 Conclusions

In this paper, we presented two remarks on Guo and Nixon’s mutual information criterion. In the first remark we explain the limitations of Guo and Nixon’s criterion and show (using Proposition 1) that the limitations arise from the higher order independence assumption made on the joint probability distribution of features. In the second remark, we discuss alternative options that can be exercised to circumvent the limitations. Our remarks intend to complement the merits of Guo and Nixon’s criterion discussed in their paper [1].

References

- [1] B. Guo and M. S. Nixon, “Gait Feature Subset Selection by Mutual Information,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 39, No. 1, January 2009.
- [2] T. Cover and J. A. Thomas, “Elements of Information Theory,” Wiley Series in Telecommunications, 1999.
- [3] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, August 2005.
- [4] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” *IEEE Transactions on Neural Networks*, Vol. 5, No. 4, July 1994.
- [5] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, New York: John Wiley & Sons, 2001.
- [6] N. Kwak and C. Choi, “Input Feature Selection by Mutual Information Based on Parzen Window,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No.12, pp.1667-1671,2002.
- [7] J. Beirlant, E.J. Dudewicz, L. Gyrfi, and E. van der Meulen, “Nonparametric Entropy Estimation: An Overview,” *International Journal of Mathematical and Statistical Sciences*, Vol. 6, No. 1, pp. 17-39, June 1997.
- [8] A. Hero, B. Ma, O. Michel, and J. Gorman, “Applications of Entropic Spanning Graphs,” *IEEE Signal Process Magazine*, Vol. 19, No. 5, pp. 85-95, 2002.
- [9] B. Bonev, F. Escolano, and M. Cazorla, “Feature Selection, Mutual Information, and the Classification of High-dimensional Patterns,” *Pattern Analysis and Applications*, Vol. 11, pp. 309-319, 2008.