Pattern Recognition Letters 101 (2018) 21–28

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A binning formula of bi-histogram for joint entropy estimation using mean square error minimization

Abdenour Hacine-Gharbi[a], Philippe Ravier[b,*]

[a] LMSE laboratory, University of Bordj Bou Arréridj, Elanasser, 34030 Bordj Bou Arréridj, Algeria
[b] PRISME laboratory, University of Orléans, INSA-CVL, 12 rue de Blois, 45067 Orléans, France

ARTICLE INFO

Article history:
Received 1 March 2017
Available online 6 November 2017

Keywords:
Histogram bin number
Bi-histogram
Marginal entropy
Joint entropy
Mutual information
Shannon entropy
Mean square error minimization
MFCC feature
Feature selection
Speech recognition
Speaker recognition

ABSTRACT

Histograms have extensively been used as a simple tool for nonparametric probability density function estimation. However, practically, the accuracy of some histogram-based derived quantities, such as the marginal entropy (ME), the joint entropy (JE), or the mutual information (MI) depends on the number of bins chosen for the histogram. In this paper, we investigate the binning problem of bi-histogram for the estimation of JE. By minimizing a theoretical mean square error (MSE) of JE estimation, we derive a new formula for the optimal number of bins of bi-histogram for continuous random variables. This novel JE estimation has been used in the MI estimation to avoid the error accumulation of joint MI between the class variable and feature subset in the feature selection. In a synthetic Gaussian feature selection problem, only the proposed method permits to retrieve the exact number of relevant features that explain the class variable when compared to a concurrent univariate estimator based on binning formula that has been proposed for ME estimation. In speech and speaker recognition applications, the proposed method permits to select a limited number of features which guaranties approximately the same or an even better recognition rate than using the total number of features.

1. Introduction

The entropy is a well-known measure of uncertainty information contained in one random variable X. When two variables X and Y are considered, the joint entropy (JE) represents uncertainty of the joint variables. Many factors may affect the JE among them the dependency between both variables. The higher the dependency between joint variables, the lower the uncertainty will be and, thus, the lower the entropy will be. When the variables are independent, uncertainty will be the highest, producing the highest entropy. This is the reason why JE is included in the definition of many information measure tools such as the dependency, distance or similarity when evaluated between two variables. In [1] and [2], several (dis)similarity measures for image registration are described as JE, mutual information (MI), normalized mutual information, symmetric uncertainty coefficient and exclusive f-information. All these measures need to compute the marginal entropy (ME) and JE between two discrete valued random variables. The JE has been used in pattern recognition [3] as dissimilarity measure for images registration [4]. The authors in [5] used MI to initialize the weights in sigmoidal feedforward neural networks. The MI between input and output variables of the neural networks were computed thanks to JE measures for evaluating useful information contained in the input variables. In all these studies using JE measures, the estimation step is essential and has to be carefully achieved. This is particularly true when the variables are continuous because the computation of JE from the data requires the integration of the joint probability density function (pdf) which is a very difficult task. Some methods introduce functions for the density estimation, such as Parzen windows [6] or Gaussian mixture models [7]. However, these methods involve an appropriate tuning of many parameter values in the selected function, which is an important issue in the estimation quality. Other common methods divide the continuous data space into several discrete partitions in order to perform discrete histogram estimation. Thus, the JE can be calculated using the definitions for discrete cases. Nevertheless, the estimation performance can be degraded as a result of large errors that are transmitted to the JE estimation due to the partitioning and pdf estimation procedures, when the number of data is limited.

The most attractive histogram partitioning is the adaptive one [8] but the computation cost dramatically increases with the number of data. So we focus in this study on the uniform partitioning because of its computational efficiency [9] which only necessitates

* Corresponding author.
E-mail address: philippe.ravier@univ-orleans.fr (P. Ravier).

https://doi.org/10.1016/j.patrec.2017.11.007
0167-8655/

one parameter to tune for histogram estimation (or two parameters for bi-dimensional histogram estimation).

In the case of a uniform bin partitioning, the number of bins k, or equivalently the bin width Δ, has to be defined for each variable. We proposed in [9] a new binning formula for the ME histogram-based estimation. This formula has been used in [5] and [10] to estimate respectively the ME and JE for MI computation. However, a direct application of this binning formula to the JE case is fundamentally not correct. We thus propose in this paper the bi-histogram binning derivation for the JE case which has to take into account the correlation between the two variables. The computation proposed in [5] will be used as a reference case (univariate) for a comparison with the new proposed appropriate JE estimation case development (bivariate).

This paper is organized as follows. Section 2 is divided into two parts. First, a general background about histogram-based estimation of JE is proposed. Second, the procedure for the optimal number of bins which produces a minimum mean square error of the histogram-based JE estimation is derived. Section 3 presents simulations that validate and show the superiority of the proposed approach. First, an experimental performance study of the new estimator is detailed. Second, the method is approved in a simulated feature selection problem. Section 4 gives the experimental results of the method for a speaker and phoneme recognition task using the logatome OLLO database [11]. Section 5 concludes the paper.

2. Method

This part is devoted to the JE definitions in the continuous and discrete cases followed by the procedure proposed to derive the appropriate number of bins. This number is obtained by following the same mean square error (MSE) minimization procedure as described in [12].

2.1. General background

Let us first recall that the entropy of a continuous random variable X, with pdf $f_X(x)$, is defined as [13]:

$$\mathrm{H}(X) = -\int_{-\infty}^{+\infty} f_X(x)\log f_X(x)\,dx \tag{1}$$

In this paper, we consider the nat measure of the entropy expressed with the natural logarithm. When two continuous random variables X and Y are considered at the same time, the entropy is a JE, with a joint pdf $f_{XY}(x, y)$, and the definition becomes [13]:

$$\mathrm{H}(X,Y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log f_{X,Y}(x,y)\,dx\,dy \tag{2}$$

Both definitions are useful for introducing the MI that quantifies the common information which is shared between the two variables X and Y. A common definition of MI writes [13]:

$$\mathrm{MI}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \tag{3}$$

This formula can be rewritten as:

$$\mathrm{MI}(X;Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)dx\,dy \tag{4}$$

We now focus on the discretization of Eq. (2). For a histogram-based JE estimation, the x-axis is divided into k_X equally sized Δ_X segments with position i and the y-axis is divided into k_Y equally sized segments with position j. The discrete approximation of (2) writes:

$$\mathrm{H}(X,Y) \approx -\sum_{i=1}^{k_X}\sum_{j=1}^{k_Y} f_{X,Y}(x_i,y_j)\log f_{X,Y}(x_i,y_j)\,\Delta_X\Delta_Y \tag{5}$$

If we now consider the probability p_{ij} of observing a sample within an area of size $(\Delta_X \times \Delta_Y)$ located around the position (i, j) in the discretization grid, an approximation of p_{ij} writes $p_{ij} \approx f_{X,Y}(x_i, x_j)\Delta_X\Delta_Y$. An estimation of this probability is classically obtained as $\hat{p}_{ij} = \frac{k_{ij}}{N}$ by counting the number of samples k_{ij} effectively falling into this area divided by the total number of samples N. So Eq. (5) results in the estimator of $\mathrm{H}(X, Y)$ as:

$$\hat{\mathrm{H}}(X,Y) = -\sum_{i=1}^{k_X}\sum_{j=1}^{k_Y}\left(\frac{k_{ij}}{N}\right)\log\left(\frac{k_{ij}}{N}\right) + \log\left(\Delta_X\Delta_Y\right) \tag{6}$$

2.2. Minimum MSE histogram-based JE estimator

In the case of arbitrary mean Gaussian distributed variables X and Y, with standard deviation σ_X and σ_Y respectively, and correlation ρ, the bias of the histogram-based JE estimation approximately writes [14]:

$$\mathrm{Bias}\left\{\hat{\mathrm{H}}(X,Y)\right\} \approx \frac{1}{24\left(1-\rho^2\right)}\left(\left(\frac{\Delta_X}{\sigma_X}\right)^2 + \left(\frac{\Delta_Y}{\sigma_Y}\right)^2\right) - \left(\frac{k_X.k_Y - 1}{2.N}\right) \tag{7}$$

The variance of the estimator approximately writes [14]:

$$\mathrm{var}\left\{\hat{\mathrm{H}}(X,Y)\right\} \approx \frac{1}{N} \tag{8}$$

The MSE of $\hat{\mathrm{H}}(X, Y)$ estimated by Eq. (6) can therefore be expressed as:

$$\mathrm{MSE}\left\{\hat{\mathrm{H}}(X,Y)\right\} \approx \frac{1}{N} + \left\{\frac{1}{24\left(1-\rho^2\right)}\left(\left(\frac{\Delta_X}{\sigma_X}\right)^2 + \left(\frac{\Delta_Y}{\sigma_Y}\right)^2\right) - \left(\frac{k_X.k_Y - 1}{2.N}\right)\right\}^2 \tag{9}$$

The purpose now is to search for the number of bins k_X and k_Y that minimize Eq. (9). In practice, the variables X and Y are of limited extension A_x and A_y that can be measured as

$A_x = \max(x) - \min(x)$ and $A_y = \max(y) - \min(y)$, where the $\max(.)$ and $\min(.)$ operators return the maximum and minimum values of the available observed data for the variables X and Y, respectively. A uniform partitioning implies that:

$$k_X\Delta_X = A_x \quad \text{and} \quad k_Y\Delta_Y = A_y \tag{10}$$

Another way to measure the extent of a random variable makes use of the standard deviation as:

$$\alpha_X\sigma_X = A_x \quad \text{and} \quad \alpha_Y\sigma_Y = A_y \tag{11}$$

where α_X and α_Y are unknown constant values depending on the distribution (typically $\alpha_X = 6$ when considering the Gaussian distribution which is assumed in the previous derivations). Eqs. (10) and (11) produce the following equalities:

$$\frac{\Delta_X}{\sigma_X} = \frac{\alpha_X}{k_X} \quad \text{and} \quad \frac{\Delta_Y}{\sigma_Y} = \frac{\alpha_Y}{k_Y} \tag{12}$$

By noting that the number of bins for Y is $k_Y = \beta\, k_X$, the issue is now to find the optimal number $k_{opt} = k_X = k_Y/\beta$ that can be obtained by solving the minimization equation:

$$k_{opt} = \arg\min_k\left[\mathrm{MSE}\left\{\hat{\mathrm{H}}(X,Y)\right\}\right]. \tag{13}$$

Since the variance of $\hat{\mathrm{H}}(X, Y)$ is independent of k, Eq. (13) can be reduced to:

$$k_{opt} = \arg\min_k\left[\mathrm{Bias}\left\{\hat{\mathrm{H}}(X,Y)\right\}^2\right]. \tag{14}$$