Pesquisa Operacional (2011) 31(3): 499-519
© 2011 Brazilian Operations Research Society
Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope

RÉNYI ENTROPY AND CAUCHY-SCHWARTZ MUTUAL INFORMATION APPLIED TO MIFS-U VARIABLE SELECTION ALGORITHM: A COMPARATIVE STUDY

Leonardo Barroso Gonçalves[1*] and José Leonardo Ribeiro Macrini[2]

Received June 27, 2009 / Accepted March 17, 2011

ABSTRACT. This paper approaches the algorithm of selection of variables named MIFS-U and presents an alternative method for estimating entropy and mutual information, "measures" that constitute the base of this selection algorithm. This method has, for foundation, the Cauchy-Schwartz quadratic mutual information and the Rényi quadratic entropy, combined, in the case of continuous variables, with Parzen Window density estimation. Experiments were accomplished with public domain data, being such method compared with the original MIFS-U algorithm, broadly used, that adopts the Shannon entropy definition and makes use, in the case of continuous variables, of the histogram density estimator. The results show small variations between the two methods, what suggest a future investigation using a classifier, such as Neural Networks, to qualitatively evaluate these results, in the light of the final objective which is greater accuracy of classification.

Keywords: variable selection, MIFS-U, entropy, mutual information, Shannon, Rényi, Parzen Window, Information-Theoretic Learning, ITL.

1 INTRODUCTION

Variable selection has a fundamental importance in classification systems, such as Neural Networks (Agrawal, Imielinski & Swami, 1993; Battiti, 1994; Joliffe, 1986). In this paper, the Mutual Information Variable Selector under Uniform Information Distribution (MIFS-U) is focused (Kwak & Choi, 2002). The objective of this algorithm is to select variables that are relevant for the output variable and at the same time reduce the redundancy among input variables. It as the name indicates is based on concepts of Information Theory, namely, entropy and mutual information (Cover & Thomas, 2006). When the variables involved are discrete, the computation of entropy and mutual information, based on the Shannon definition, is simple

*Corresponding author
[1] Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. E-mail: leogonet@ig.com.br
[2] Departamento de Ciências Econômicas e Exatas, Universidade Federal Rural do Rio de Janeiro, Três Rios, RJ, Brasil. E-mail: macrini@centroin.com.br

and direct, since the joint and marginal distributions can be estimated simply by counting the samples. However, when at least one of the variables in question is continuous, the computation that involves integration becomes difficult due to the limited number of samples. A solution is usually to insert the discretization of the data as a step of pre-processing, and to estimate the unknown density by the histogram. Not always, however, the discretization is made clearly and adequately. This paper shows a alternative method based on the Cauchy-Schwartz quadratic mutual information and the Rényi quadratic entropy, this combined with the Parzen Window density estimator (Silverman, 1986), and in this way the computations become direct without need of a pre-processing step. Initially, this paper shows a introduction to information theory based on the Shannon and the Rényi entropies and additionally shows the Cauchy-Schwartz mutual information, concept used in Information-Theoretic Learning (ITL) (Príncipe, 2000; Príncipe, Fisher & Xu, 1998). Next, the MIFS-U variable selector and the estimation methods of entropy and mutual information are shown. Finally, both methods are applied in datasets and the results are compared in order to obtain an initial notion of the performance of the proposed method.

2 INFORMATION THEORY

Information theory was developed by Shannon in communication engineering applications in the early 1940s. This theory, due to its innovative character and mathematical elegance, had great impact not only in engineering but also in several areas such as statistics and economy.

This section summarizes, in a descriptive way, theoretical foundations of information theory. For demonstrations and explanations on the subject matter, see, in particular, reference Cover & Thomas (2006).

2.1 Shannon Entropy and Shannon Mutual Information

The uncertainty characterizes the information gain that the occurrence of an event can cause. Therefore, such uncertainty can be translated into the probability of occurrence of its event. An event whose occurrence is right doesn't bring any increment of information, because the whole information is already contained in its occurrence certainty. In this way, it can be said that the determination of the amount of information produced by the occurrence of an event is determined by the amount of "surprise" that such occurrence brings.

The entropy in information theory corresponds therefore to the probabilistic uncertainty associated with a probability distribution.

Definition 1 – The *Shannon entropy* $H(X)$ of a discrete random variable X, with a probability mass function $f_X(x)$, $x \in \mathrm{X}$ (the domain set of the variable), is defined by

$$H(X) = -\sum_{x \in \mathrm{X}} f_X(x)\ \log f_X(x) \tag{1}$$

It results from the own definition that $H(X) \geq 0$. It is commom to denote the above quantity by $H(f_X)$.

Definition 2 – The *(Shannon) joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint probability distribution f_{xy} is defined as

$$H(X, Y) = -\sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} f_{xy}(x, y)\ \log f_{xy}(x, y) \tag{2}$$

Definition 3 – The *(Shannon) conditional entropy* $H(Y|X)$ (that is, of Y given the knowledge of X) is defined as

$$H(Y|X) = -\sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} f_{xy}(x, y) \log f_{y|x}(y|x) \tag{3}$$

Definition 4 – The *Kullback-Leibler relative entropy* (or, *(asymmetrical) divergence*) between two probability mass functions f and g is defined as

$$D_{KL}(f \| g) = \sum_{x \in \mathrm{X}} f(x) \log \frac{f(x)}{g(x)} = E_f \left(\log \frac{f(x)}{g(x)} \right) \tag{4}$$

$D_{KL}(f \| g) \geq 0$ with equality if and only if $f(x) = g(x)$, for every $x \in \mathrm{X}$.

The Kullback-Leibler relative entropy (or, divergence) is a similarity measure between strictly positive functions, it is also referred as “distance” between distributions, however, it is not a true distance since it is not symmetric and does not satisfy the triangle inequality. It is very used in the comparison between two functions. In that case, the function g represents the reference function. The Kullback-Leibler divergence is intimately related with the Shannon entropy.

Definition 5 – The *(Shannon) mutual information* $I(X; Y)$ between two discrete random variables X and Y, with a joint probability mass function $f_{xy}(x, y)$ and marginal probability mass functions $f_x(x)$ and $f_y(y)$ is given by the relative entropy between the joint distribution and the product of the marginal distributions:

$$I(X; Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} f_{xy}(x, y) \log \frac{f_{xy}(x, y)}{f_x(x) f_y(y)} = D_{KL}(f_{xy}(x, y) \| f_x(x) f_y(y)) \geq 0 \tag{5}$$

with equality if and only if $f_{xy}(x, y) = f_x(x) f_y(y)$ (that is, if X and Y are independent).

Starting from Eq. (5), it can be said that the mutual information is a measure of statistical independence. The higher the mutual information, the stronger the association between the variables.

Next, the concept of differential entropy is introduced that it is the entropy of a continuous random variable. The differential entropy is similar in many forms to the entropy of a discrete random variable, but there are important differences, and it is therefore necessary to pay attention to the usage of that concept. Note that unlike the discrete case, the differential entropy can be negative but, as it will be seen, the differential version of the mutual information will always be non-negative.

Definition 6 – The *Shannon differential entropy* $h(X)$ of a continuous random variable X, with a probability density function $f_x(x)$, $x \in \mathrm{X}$ (the support set of the variable), is defined by

$$h(X) = -\int_{\mathrm{X}} f_x(x) \log f_x(x)\, dx \tag{6}$$

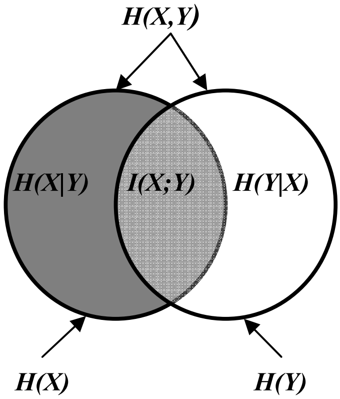

Figure 1 – The relation between the entropy and the mutual information.

Definition 7 – The *(Shannon) joint differential entropy* $h(X, Y)$ of a pair of continuous random variables (X, Y), with joint density f_{xy}, is defined as

$$h(X, Y) = -\int_{\mathrm{Y}} \int_{\mathrm{X}} f_{xy}(x, y) \log f_{xy}(x, y)\, dx\, dy \tag{7}$$

Definition 8 – The *(Shannon) conditional differential entropy* $h(X|Y)$ is defined as

$$h(X|Y) = -\int_{\mathrm{Y}} \int_{\mathrm{X}} f_{xy}(x, y) \log f_{x|y}(x|y)\, dx\, dy \tag{8}$$

Definition 9 – The *Kullback-Leibler relative entropy* (or, *(asymmetrical) divergence*) between two densities f and g is defined as

$$D_{KL}(f\|g) = \int_{\mathrm{X}} f(x) \log \frac{f(x)}{g(x)}\, dx \ = E_f\left(\log \frac{f(x)}{g(x)}\right) \tag{9}$$

$D_{KL}(f\|g) \geq 0$ with equality if and only if $f = g$ *almost everywhere*.

Definition 10 – The *(Shannon) mutual information* $I(X; Y)$ between two continuous random variables X and Y, with joint density f_{XY} and marginal densities f_X and f_Y is defined by

$$I(X; Y) = \int_{\mathrm{Y}} \int_{\mathrm{X}} f(x, y) \log \frac{f_{xy}(x, y)}{f_x(x) f_y(y)}\, dx\, dy = D_{KL}(f_{xy}(x, y)\|f_x(x) f_y(y)) \geq 0 \tag{10}$$

Although, unlike the entropy for discrete random variables, the differential entropy cannot be interpreted as a randomness (or uncertainty) measure, the mutual information has the same interpretation as in the discrete case.

2.2 Rényi Entropy and Rényi Mutual Information

Definition 11 – The *order α Rényi entropy* $H_{R_\alpha}(X)$ of a discrete random variable X, with a probability mass function $f_x(x)$, $x \in \mathrm{X}$, is defined as

$$H_{R_\alpha}(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathrm{X}} f_x^{\alpha}(x), \ \text{ for } \alpha > 0 \text{ and } \alpha \neq 1 \tag{11}$$

The Shannon entropy appears as a special case of the Rényi entropy by taking the limit of it as $\alpha \to 1$.

Of particular interest is Rényi's entropy of order two, which is called the *Rényi quadratic entropy*.

Definition 12 – The *order α differential Rényi entropy* $h_{R_\alpha}(X)$ of a continuous random variable X, with a probability density function $f_x(x)$, is defined as

$$h_{R_\alpha}(X) = \frac{1}{1-\alpha} \log \left[\int_{\mathrm{X}} f_x^\alpha(x) dx \right] \tag{12}$$

Again, of particular interest is when $\alpha = 2$, and it is denoted *differential Rényi quadratic entropy*.

Definition 13 – The *Rényi relative entropy* (or, *(asymmetrical) divergence*) *of order* α between two probability mass functions f and g is defined as

$$D_{R_\alpha}(f \| g) = \frac{1}{\alpha-1} \log \sum_{x \in \mathrm{X}} g(x) \left(\frac{f(x)}{g(x)} \right)^\alpha = \frac{1}{\alpha-1} \log \sum_{x \in \mathrm{X}} f^\alpha(x)\, g^{1-\alpha}(x), \tag{13}$$

$$\text{for } \alpha > 0 \text{ and } \alpha \neq 1$$

$D_{R_\alpha}(f \| g) \geq 0$ with equality if and only if $f(x) = g(x)$, for every $x \in \mathrm{X}$. Note that the Kullback-Leibler divergence is obtained in the limit as $\alpha \to 1$.

Definition 14 – The *Rényi relative entropy* (or, *(asymmetrical) divergence*) *of order* α between two densities f and g (Neemuchwala, 2005) is defined as

$$D_{R_\alpha}(f \| g) = \frac{1}{\alpha-1} \log \int_{\mathrm{X}} g(x) \left(\frac{f(x)}{g(x)} \right)^\alpha dx = \frac{1}{\alpha-1} \log \int_{\mathrm{X}} f^\alpha(x)\, g^{1-\alpha}(x)\, dx, \tag{14}$$

$$\text{for } \alpha > 0 \text{ and } \alpha \neq 1$$

$D_{R_\alpha}(f \| g) \geq 0$ with equality if and only if $f = g$ *almost everywhere.*

The Rényi divergence may also be used as a measure of mutual information between random variables, by considering the divergence between the joint distribution and the product of marginal distributions, according to the following definitions which are based only on the quadratic divergence (of order 2), being simply represented by $I_R(X; Y)$.

Definition 15 – The *Rényi mutual information* $I_R(X; Y)$ between two discrete random variables X and Y, with a joint probability mass function $f_{xy}(x, y)$ and marginal probability mass functions $f_x(x)$ and $f_y(y)$ is given by the relative entropy between the joint distribution and the product of the marginal distributions:

$$I_R(X; Y) = \log \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} \frac{f_{xy}^2(x, y)}{f_x(x) f_y(y)} = D_{R_2}(f_{xy}(x, y) \| f_x(x) f_y(y)) \geq 0 \tag{15}$$

with equality if and only if $f_{xy}(x, y) = f_x(x) f_y(y)$ (that is, if X and Y are independent).

Definition 16 – The *Rényi mutual information* $I_R(X;Y)$ between two continuous random variables X and Y, with joint density $f_{xy}(x,y)$ and marginal densities $f_x(x)$ and $f_y(y)$ is defined by

$$I_R(X;Y) = \log \int_{\mathrm{Y}} \int_{\mathrm{X}} \frac{f_{xy}^2(x,y)}{f_x(x) f_y(y)}\, dx\, dy = D_{R_2}(f_{xy}(x,y) \| f_x(x) f_y(y)) \geq 0 \tag{16}$$

with equality if and only if $f_{xy} = f_x f_y$ *almost everywhere* (that is, if X and Y are independent).

The Rényi mutual information, unlike the Shannon mutual information, can not be expressed in terms of Rényi entropies (Jenssen, 2005). However, the Cauchy-Schwartz mutual information, which is introduced in the next section, can be expressed, as will be seen, by Rényi quadratic entropy.

2.3 Cauchy-Schwartz Mutual Information

Principe *et al.* (2000) defined a measure of divergence between probability density functions (or probability mass functions) based on the Cauchy-Schwartz inequality between vectors.

Definition 17 – The *Cauchy-Schwartz (simetrical) divergence* between two probability mass functions $f(x)$ and $g(x)$ is defined by

$$D_{CS}(f \| g) = -\log \frac{\sum\limits_{x \in \mathrm{X}} f(x) g(x)}{\sqrt{\left(\sum\limits_{x \in \mathrm{X}} f^2(x)\right)\left(\sum\limits_{x \in \mathrm{X}} g^2(x)\right)}} \tag{17}$$

$D_{CS}(f \| g) \geq 0$ with equality if and only if $f(x) = g(x)$, for every $x \in \mathrm{X}$.

Developing the previous equation, one gets

$$D_{CS}(f \| g) = -\log \sum_{x \in \mathrm{X}} f(x) g(x) - \frac{1}{2}\left(-\log \sum_{x \in \mathrm{X}} f^2(x)\right) - \frac{1}{2}\left(-\log \sum_{x \in \mathrm{X}} g^2(x)\right) \tag{18}$$

Expressing the second member of Eq. (18) through the Rényi entropy, the following is obtained:

$$D_{CS}(f \| g) = h_{R_2}(f \times g) - \frac{1}{2} h_{R_2}(f) - \frac{1}{2} h_{R_2}(g) \tag{19}$$

where

- $h_{R_2}(f)$ is the Rényi quadratic entropy with respect to f.
- $h_{R_2}(g)$ is the Rényi quadratic entropy with respect to g.
- $h_{R_2}(f \times g)$ can be interpreted as the *cross-entropy* between f and g.

Definition 18 – The *Cauchy-Schwartz (simetrical) divergence* between two densities f and g is defined by

$$D_{CS}(f \| g) = -\log \frac{\int_{\mathrm{X}} f(x) g(x) dx}{\sqrt{\left(\int_{\mathrm{X}} f^2(x) dx\right)\left(\int_{\mathrm{X}} g^2(x) dx\right)}} \tag{20}$$

$D_{CS}(f\|g) \geq 0$, with equality if and only if $f = g$ *almost everywhere*, and the integrals involved are all quadratic forms of probability density functions.

In a similar way, Eq. (19) is also obtained by developing the previous equation.

Definition 19 – The *Cauchy-Schwartz mutual information* $I_{CS}(X;Y)$ between two discrete random variables X and Y, with a joint probability mass function $f_{xy}(x,y)$ and marginal probability mass functions $f_x(x)$ and $f_y(y)$ is given by the divergence between the joint distribution and the product of the marginal distributions:

$$I_{CS}(X;Y) = h_{R_2}(f_{XY} \times f_X f_Y) - \frac{1}{2}h_{R_2}(f_{XY}) - \frac{1}{2}h_{R_2}(f_X f_Y) = D_{CS}(f_{xy}\|f_x f_y) \geq 0 \qquad (21)$$

with equality if and only if $f_{xy}(x,y) = f_x(x)f_y(y)$ (that is, if X and Y are independent).

Definition 20 – The *Cauchy-Schwartz mutual information* $I_{CS}(X;Y)$ between two continuous random variables X and Y, with joint density $f_{xy}(x,y)$ and marginal densities $f_x(x)$ and $f_y(y)$ is given by Eq. (21), with equality if and only if $f_{xy} = f_x f_y$ *almost everywhere* (that is, if X and Y are independent).

3 MUTUAL INFORMATION VARIABLE SELECTION UNDER UNIFORM INFORMATION DISTRIBUTION (MIFS-U)

Input variable selection plays an important role in classifying systems such as neural networks (NNs). A input variable can be classified as relevant, irrelevant or redundant and from the viewpoint of managing a dataset which can be huge, reducing the number of variables by selecting only the relevant ones is desirable. In doing so, higher performances with lower computational effort is expected (Kwak & Choi, 2002).

Hosmer & Lemeshow (1989) highlight the importance of variable selection, because with a smaller number of variables, the model tends to be more generalizable and robust.

Problems of variable selection has been tackled by several researchers such as Battiti (1994), Joliffe (1986) and Agrawal *et al.* (1993). One of the most popular methods for dealing with this problem is the principal component analysis (PCA) method (Joliffe, 1986). However, when the maintenance of the original variables is wanted, this method is not adequate.

The algorithm approached in this paper, that is the MIFS-U – Mutual Information Feature (Variable) Selector under Uniform Information Distribution – was presented by Kwak & Choi (2002), with the objective of overcoming the limitation of variable selector proposed by Battiti (1994), producing better performance of the variable selection procedure. Such algorithm can be used in any classifying systems for its simplicity whatever the learning algorithm may be. But the performance can be degraded as a result of errors in estimating the mutual information.

3.1 The FRn-k Problem and the Ideal Selection Algorithm

In the process of selecting input variables, it is desirable to reduce the number of variables by excluding irrelevant or redudant variables among the ones. This concept is formalized as selecting the most relevant k variables from a set of n variables and Battiti (1994) named it as "feature reduction – FR" problem. Such process is described as follows:

[FRn – k]: Given an initial set of n variables, find the subset with $k < n$ variables that is "maximally informative" about the class (output variable). The problem of selecting input variables can be solved by computing the mutual information (MI) between input variables and output classes. If the mutual information between input variables and output classes could be exactly obtained, the FRn – k problem could be reformulated as follows:

[FRn – k]: Given an initial set F with n variables and the output variable D, find the subset $S \subset F$ with k variables that minimizes $H(D|S)$, that is, that maximizes the mutual information $I(D; S)$. The selection method here adopted is known as "greedy selection". In this method, from the empty set of selected variables, the best input variable of the current state is added one by one. This *ideal selection algorithm* using mutual information is realized as follows:

1) (Initialization) set $F \leftarrow$ "initial set of n variables", $S \leftarrow$ "empty set."
2) (Computation of the MI with the output class), $\forall \phi_i \in F$, compute $I(D; \phi_i)$.
3) (Selection of the first variable) find the variable that maximizes $I(D; \phi_i)$, set $F \leftarrow F \backslash \{\phi_i\}$, $S \leftarrow \{\phi_i\}$.
4) (Greedy selection) repeat until desired number of variables are selected:
    a) (Computation of the joint MI between variables), $\forall \phi_i \in F$, compute $I(D; \phi_i, S)$.
    b) (Selection of the next variable) choose the variable $\phi_i \in F$ that maximizes $I(D; \phi_i, S)$, and set $F \leftarrow F \backslash \{\phi_i\}$, $S \leftarrow \{\phi_i\}$.
5) Output the set S containing the selected variables.

In practice, the realization of this algorithm is unviable due to the high dimensionality of the vector of variables in the computation of $I(D; \phi_i, S)$, since the objective is to select $k(k < n)$ variables, and therefore the vector S (composed of the variables already selected), reaches dimension $(k - 1)$.

3.2 The MIFS-U Variable Selector

The ideal algorithm (Battiti, 1994) tries to maximize $I(D; \phi_i, \phi_s)$ (area II, III and IV in Fig. 2) and, according to Kwak & Choi (2002), this can be rewritten as

$$I(D; \phi_i, \phi_s) = I(D; \phi_s) + I(D; \phi_i|\phi_s), \tag{22}$$

where $I(D; \phi_i|\phi_s)$ represents the remaining mutual information between the output class D and the variable ϕ_i for a given ϕ_s. This is shown as area III in Figure 2, whereas the area II plus area IV represents $I(D; \phi_s)$. Since $I(D; \phi_s)$ is common for all the candidate variables to be selected in the ideal algorithm, there is no need to compute this. So the ideal algorithm tries to find the variable that maximizes $I(D;\phi_i|\phi_s)$ (area III). However, calculating $I(D;\phi_i|\phi_s)$ requires as much work as calculating $I(D;\phi_i;\phi_s)$. So $I(D;\phi_i|\phi_s)$ is approximately computed with $I(\phi_i;\phi_s)$ and $I(D;\phi_i)$, which are relatively easy to calculate. The conditional mutual information can be represented as

$$I(D;\phi_i|\phi_s) = I(D;\phi_i) - \{I(\phi_i;\phi_s) - I(\phi_i;\phi_s|D)\} \tag{23}$$

where $I(\phi_i;\phi_s)$ corresponds to arera I and IV, and $I(\phi_i;\phi_s|D)$ corresponds to area I. So the term $I(\phi_i;\phi_s) - I(\phi_i;\phi_s|D)$ corresponds to area IV. The term $I(\phi_i;\phi_s|D)$ means the mutual information between the already selected variable ϕ_s and the candidate variable ϕ_i for a given class D.

If conditioning by the class D does not change the ratio of the entropy of ϕ_s and the mutual information between ϕ_i and ϕ_s, that is, if the following relations holds (condition of the algorithm):

$$\frac{H(\phi_s|D)}{H(\phi_s)} = \frac{I(\phi_i;\phi_s|D)}{I(\phi_i;\phi_s)} \tag{24}$$

$I(\phi_i;\phi_s|D)$ can be represented as

$$I(\phi_i;\phi_S|D) = \frac{H(\phi_s|D)}{H(\phi_s)}\, I(\phi_i;\phi_s) \tag{25}$$

Using the equation above and Eq. (23), the following is obtained:

$$I(D;\phi_i|\phi_s) = I(D;\phi_i) - \frac{I(D;\phi_s)}{H(\phi_s)}\, I(\phi_i;\phi_s). \tag{26}$$

Assuming that each region in Figure 2 corresponds to its corresponding information, the condition presented in Eq. (24) is hard to satisfied when information is concentrated on one of the following regions: $H(\phi_s|\phi_i;D)$, $I(\phi_s;\phi_i|D)$, $I(D;\phi_s|\phi_i)$ or $I(D;\phi_s;\phi_i)$. It is more likely that condition (24) hods when information is distributed uniformly throughout the region of $H(\phi_s)$ in Figure 2. Because of this, the algorithm is simply called the MIFS-U algorithm.

Then the revised step 4 of the ideal selection algorithm takes the following form:

4) (Greedy selection) repeat until desired number of variables are selected:

   a) (Computation of entropy) $\forall\phi_s \in S$, compute $H(\phi_s)$, if is not already available.

   b) (Computation of the MI between variables), for all couples of variables (ϕ_i,ϕ_s) with $\phi_i \in F$ and $\phi_s \in S$, compute $I(\phi_i;\phi_s)$, if it is not yet available.

   c) (Selection of the next variable) choose a variable $\phi_i \in F$ that maximizes $I(D;\phi_i) - \beta\sum_{\phi_s\in S}(I(D;\phi_s)/H(\phi_s))\ I(\phi_i;\phi_s)$ and set $F \leftarrow F\backslash\{\phi_i\}$, $S \leftarrow \{\phi_i\}$.

Parameter β offers flexibility to the algorithm as in the MIFS. If $\beta = 0$, the mutual information between input variables is not considered and the algorithm chooses input variables in the order of

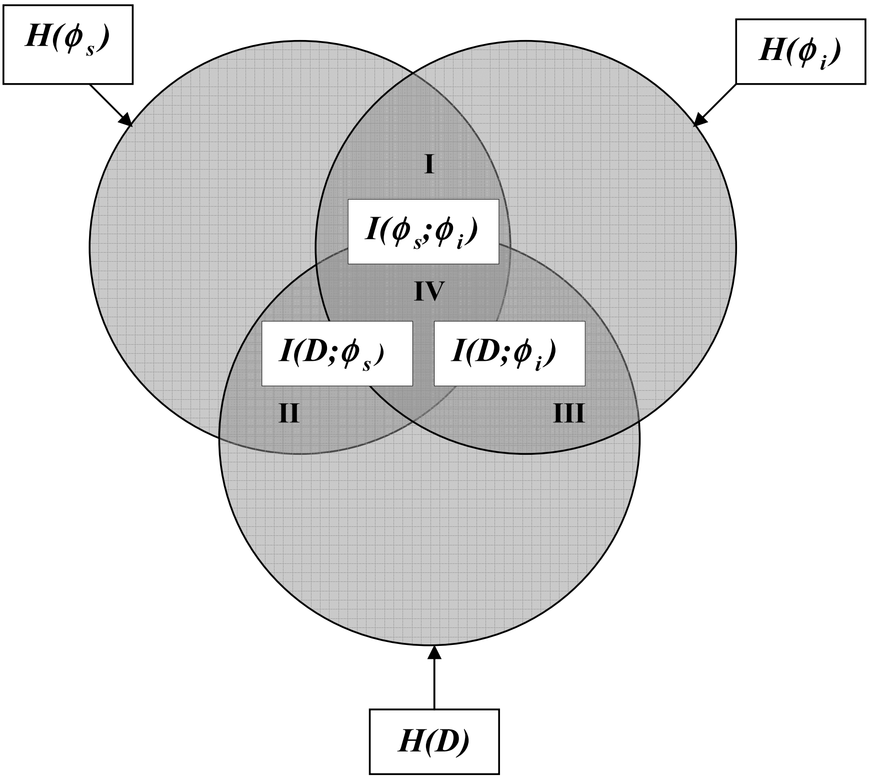

Figure 2 – The relation between input variables and output classes.

the mutual information with the output. As β grows aumenta, it excludes the redundant variables more efficiently. In general β can be taken as 1 (Breiman *et al.*, 1984). In this case there is a balance in terms of weight between the redundancy of the candidate variable and the mutual information between this variable and the output. So, for all the experiments in this paper, $\beta = 1$ is adopted.

Kwak & Choi (2002) point out that the MIFS-U algorithm can be applied to large problems without excessive computational efforts.

4 ESTIMATION METHODS OF ENTROPY AND MUTUAL INFORMATION

The estimation of entropy and mutual information, involving only discrete random variables, is simple, with direct application of the Shannon definition. However, when one of the involved variables is continuous, it is necessary to apply a density estimation method. One of the simplest and most widely-used is the Histogram. From now on this method forward described will be called Shannon/Histogram Method. The second method, which is presented as an alternative, is based on the Rényi quadratic entropy, combined with Parzen Window density estimation, and on the Cauchy-Schwartz Mutual Information. In this way the computations become direct without need of a pre-processing step. From now on this method will be called Cauchy-Schwartz/Parzen-Rosenblatt Method.

4.1 Shannon/Histogram Method

In the case of continuous variables, to avoid adopting a parametric model for the unknown density, a common solution is to apply non-parametric density estimation methods. The oldest and the most widely used density estimator is the histogram (Silverman, 1986). In this paper, the relative frequency histogram is actually used, not the density histogram, where the only difference is that the latter is normalized to integrate to 1 (Scott, 1992).

As all the continuous variables are normalized in the interval $[-1, 1]$, the interval is simply divided into 20 subintervals of equal width $(h = 0, 1)$. Each subinterval is interpreted as a class and each computed relative frequency is taken as a probability. In other words, a discretization – a continuous variable becomes discrete – is done. Then there are no more obstacle to the necessary computations, and the Shannon entropy definition, widely used in the literature, can be easily applied.

In order to maintain a harmonic nomenclature, a specific class of a discretized continuous variable or a (distinct) value of a discrete variable will simply be represented by x (and the set of such values or classes will be represented by X), and therefore, this distinction is no longer needed.

So the following can be written:

$$\hat{f}_x(x) = \frac{1}{n}\sum_{i=1}^{n} \xi(x_i, x) \quad , \forall x \in \mathrm{X}. \tag{27}$$

where $\xi_x(x_i)$ is the Indicator Function, that is,

$$\xi(x_i, x) = \begin{cases} 1 & , \text{ if } x_i \in x \text{ (class)} \\ 0 & , \text{ otherwise} \end{cases} \qquad \text{(Discretized Continuous Variable)}$$

or

$$\xi(x_i, x) = \begin{cases} 1 & , \text{ if } x_i = x \text{ (value)} \\ 0 & , \text{ otherwise} \end{cases} \qquad \text{(Discrete Variable)}$$

and the joint distribution of two (discrete or discretized continuous) variables is the following:

$$\hat{f}_{xy}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \xi_x(x_i)\xi_y(y_i) \; ,\forall x \in \mathrm{X},\ \forall y \in \mathrm{Y}. \tag{28}$$

Therefore, considering the discretization of continuous variables as a step of pre-processing of the data, the necessary computations for the MIFS-U algorithm are merely the following ones:

- Entropy of a discrete variable (EntD),
- Mutual information between two discrete variables (MI-DD).

EntD – Shannon Entropy of a discrete variable

$$\hat{H}(X) = -\sum_{x \in \mathrm{X}} \hat{f}_x(x) \, \log \hat{f}_x(x) \tag{29}$$

MI-DD – (Shannon) Mutual Information between two discrete variables

$$\hat{I}(X; Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} \hat{f}_{xy}(x, y) \log \frac{\hat{f}_{xy}(x, y)}{\hat{f}_x(x) \hat{f}_y(y)} \tag{30}$$

4.2 Cauchy-Schwartz/Parzen-Rosenblatt Method

In the context of variable selection in nonlinear systems, the estimation of the mutual information between variables directly from the data, where at least one of them is continuous, without hypotheses about the *priori* distribution of the data, has vital practical importance. This can be reached using the Cauchy-Schwartz divergence, which is a substitute of the Kullback-Leibler divergence, integrated with the Parzen Window estimator.

The Kullback-Leibler divergence, based on the Shannon entropy, is, in its simplicity, an usual measure of mutual information between two random variables. However, neither this nor the equivalent for the Rényi entropy can be integrated with the Parzen Window estimator (Príncipe et al., 1998). Xu *et al.* (1998) presented a method that combines the Cauchy-Schwartz Divergence with Parzen Windowing for estimating the mutual information directly from the data.

4.2.1 Parzen Window Density Estimator

According to Scott (1992), given a set of samples of f_x $\{x_1, x_2, \ldots, x_n\}$, the Kernel Density Estimator – or the Parzen Window Estimator – may be written compactly as

$$\hat{f}_x(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i, h) \tag{31}$$

where $h = h(n) > 0$ is the window width or smoothing parameter.

So $\hat{f}_x(x)$ can be seen as an "average of curves" centered at the samples.

The Kernel Function $K(\cdot)$ is usually non-negative and with unitary integral, that is, a probability density function (Silverman, 1986). Furthermore, often $K(\cdot)$ is chosen to be a symmetric and unimodal density.

For the later use in this paper, the Gaussian Kernel Function will be considered, defined below.

$$G(w, \phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\left(-\frac{\omega^2}{2\phi}\right) \tag{32}$$

that is, $G(w, \phi) \sim N(0, \phi)$, where $\phi = h^2 = \sigma^2$.

The choice of the window width affects the density estimate much more than the choice of the Kernel Function (Scott, 1992). So the choice of the Kernel Function is not crucially important. However, the Gaussian Function has a property that will be extremely advantageous in the context of this paper.

There exist several methods for selecting the window width h, each having its properties (Wand & Jones, 1995). The method here used is known as "Normal Reference Rule", and the window width is given by

$$h_{ot} = 1,06\,\sigma\, n^{-1/5} \tag{33}$$

The standard deviation σ can be estimated, starting from the data, by the sample standard deviation s or by a robust measure like the interquartile range. In this paper, a commitment solution between both estimators is used, similar to the form presented by Silverman (1986):

$$\hat{h}_{ot} = 0,9\ \min(s, \hat{I}_q)\, n^{-1/5} \tag{34}$$

In the bivariate case, using a single window width for both variables and taking the same considerations in regard to the estimation of the standard deviation in the univariate case, the following is adopted in this paper:

$$h_{ot} \approx 0,85 \min\left(\left(\frac{s_1^2 + s_2^2}{2}\right)^{-\frac{1}{2}}, \frac{\hat{I}_q^{(1)} + \hat{I}_q^{(2)}}{2}\right) n^{-\frac{1}{6}} \tag{35}$$

4.2.2 Necessary Computations for the MIFS-U Algorithm

The types of computation required for the MIFS-U algorithm are presented below, using indistinctly the notation h_{R_2} in the representation the Rényi entropy, whether differential or not.

EntD – Rényi Entropy of a discrete variable

$$\hat{h}_{R_2}(X) = -\log \sum_{x \in \mathrm{X}} \hat{f}_x^2(x) \tag{36}$$

EntC – Rényi Entropy of a continuous variable

Estimating the density through the Gaussian Kernel Function

$$\hat{f}_x(x) = \frac{1}{n}\left(\sum_{i=1}^{n} G(x - x_i, \sigma^2)\right) \tag{37}$$

and applying the property of the integration of the product of Gaussian kernels, shown below,

$$\int G\big(x - a_i, \sigma_1^2\big)\, G\big(x - a_j, \sigma_2^2\big) dx = G\big(a_i - a_j, \sigma_1^2 + \sigma_2^2\big) \tag{38}$$

the Rényi quadratic entropy is easily estimated by

$$\hat{h}_{R_2}(X) = -\log\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G\big(x_i - x_j, 2\sigma^2\big)\right] \tag{39}$$

Thus, the Rényi quadratic entropy can be estimated as a sum of local interactions, as defined by the kernel, over all pairs of samples.

MI-DD – (Cauchy-Schwartz) Mutual Information between two discrete variables

$$I_{CS}(X;Y) = h_{R_2}(f_{XY} \times f_X f_Y) - \frac{1}{2}h_{R_2}(f_{XY}) - \frac{1}{2}h_{R_2}(f_X f_Y) \tag{40}$$

where

$$\hat{h}_{R_2}(\hat{f}_{xy} \times \hat{f}_x \hat{f}_y) = -\log \sum_{\text{Y}}\sum_{\text{X}} \hat{f}_{xy}(x,y)\hat{f}_x(x)\hat{f}_y(y) \tag{41}$$

$$\hat{h}_{R\,2}(\hat{f}_{xy}) = -\log \sum_{\text{Y}}\sum_{\text{X}} \hat{f}_{xy}^2(x,y) \tag{42}$$

$$\hat{h}_{R\,2}(\hat{f}_x \hat{f}_y) = -\log \sum_{\text{Y}}\sum_{\text{X}} \hat{f}_x^2(x)\hat{f}_y^2(y) \tag{43}$$

MI-CC – (Cauchy-Schwartz) Mutual Information between two continuous variables

As seen, the entropy of a single variable is easily evaluated as interactions between pairs of samples. This concept will now be extended to mutual information between variables. Thus, in the continuous case, Eq. (40) is the second member of the equation given by

$$\begin{aligned}\hat{h}_{R\,2}(\hat{f}_{xy} \times \hat{f}_x \hat{f}_y) &= -\log \int_{\text{Y}}\int_{\text{X}} \hat{f}_{xy}(x,y)\hat{f}_x(x)\hat{f}_y(y)dx\,dy \\ &= -\log\left\{\frac{1}{n^3}\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\right)\left(\sum_{l=1}^{n} G(y_i - y_l, 2\sigma^2)\right)\right]\right\}\end{aligned} \tag{44}$$

$$\begin{aligned}\hat{h}_{R\,2}(\hat{f}_{xy}) &= -\log \int_{\text{Y}}\int_{\text{X}} \hat{f}_{xy}^2(x,y)dx\,dy \\ &= -\log\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\,G(y_i - y_j, 2\sigma^2)\right]\end{aligned} \tag{45}$$

$$\begin{aligned}\hat{h}_{R\,2}(\hat{f}_x \hat{f}_y) &= -\log \int_{\text{Y}}\int_{\text{X}} \hat{f}_x^2(x)\hat{f}_y^2(y)dx\,dy \\ &= -\log\left\{\frac{1}{n^4}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, 2\sigma^2)\right]\left[\sum_{k=1}^{n}\sum_{l=1}^{n} G(y_k - y_l, 2\sigma^2)\right]\right\}\end{aligned} \tag{46}$$

MI-DC – (Cauchy-Schwartz) Mutual Information between a discrete variable (Y) and a continuous variable (X)

Consider the following definitions:

$v =$ number of distinct values of Y in the sample.

$y_p = p$th distinct value of Y in the sample.

$n_p =$ number of samples of X related to the value y_p of Y.

Here two different notations are used for the samples of X. A sample is written with a single subscript $x_i (1 \le i \le n)$ when the identification of the Y-value related to it is irrelevant. If it is relevant, x_{ps} indicates the sample of X, with index $1 \le s \le n_p$, related to the value y_p $(1 \le p \le v)$ of Y.

$$\hat{f}_y(y_p) = \frac{n_p}{n} \qquad \sum_{p=1}^{v} n_p = n \tag{47}$$

Estimating the densities through the Gaussian Kernel Function, in the case of the continuous variable, and using the property of the integration of the product of Gaussian kernels, the entropies appearing in Eq. (40) are estimated by

$$\begin{aligned} \hat{h}_{R\,2}(\hat{f}_{xy} \times \hat{f}_x \hat{f}_y) &= -\log \sum_{p=1}^{v} \int_{\mathrm{X}} \hat{f}_{xy}(x, y_p) \hat{f}_x(x) \hat{f}_y(y_p)\, dx \\ &= -\log \frac{1}{n^3} \sum_{p=1}^{v} \left[n_p \sum_{s=1}^{n_p} \sum_{i=1}^{n} G\big(x_{ps} - x_i, 2\sigma^2\big) \right] \end{aligned} \tag{48}$$

$$\hat{h}_{R\,2}(\hat{f}_{xy}) = -\log \sum_{p=1}^{v} \int_{\mathrm{Y}} \hat{f}_{xy}^2(x_i, y)\, dy = -\log \frac{1}{n^2} \sum_{p=1}^{v} \sum_{s=1}^{n_p} \sum_{t=1}^{n_p} G\big(x_{ps} - x_{pt}, 2\sigma^2\big) \tag{49}$$

$$\begin{aligned} \hat{h}_{R\,2}(\hat{f}_x \hat{f}_y) &= -\log \sum_{p=1}^{v} \int_{\mathrm{X}} \hat{f}_x^2(x) \hat{f}_y^2(y_p)\, dx \\ &= -\log \left[\frac{1}{n^4} \left[\left(\sum_{p=1}^{v} n_p^2 \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{n} G\big(x_i - x_j, 2\sigma^2\big) \right) \right] \right] \end{aligned} \tag{50}$$

5 EXPERIMENTS

5.1 A Brief Description of the Databases

The databases were extracted from the *UCI Machine Learning Repository* (http://archive.ics.uci.edu/ml/datasets.html).

It is not in the scope of this study a specific analysis of the databases, since the use of the databases considered here has in view the mere comparison of the results regarding the selection order by the MIFS-U algorithm, considering the two estimation methods of entropy and mutual information presented in this paper.

In the subsequent table, it can be observed the following information:

- the number of complete samples considered in each database (n),
- the number of discrete and continuous variables (ignoring the output).

Table 1 – Databases.

Databases	n	Number of variables	
		Discrete	Continuous
ECHOCARDIOGRAM *Echocardiogram Data*	61	3	8
TELESCOPE *Magic gamma telescope data 2004*	554	0	10
WINE *Wine recognition data*	130	0	13
DERMATOLOGY *Dermatology Database*	171	27	1
BREAST CANCER *Wisconsin Diagnostic Breast Cancer (WDBC)*	569	0	30
HEART DISEASE *Heart Disease Database*	214	8	5

5.2 Comparison of the Methods

The comparison of the results of the selection by the MIFS-U, regarding both estimation methods of entropy and mutual information presented in this paper, is shown in following tables. The values are normalized to 1. The analysis focuses the first five selected variables. For simplification, the Shannon/Histogram and Cauchy-Schwartz/Parzen-Rosenblatt Methods will be respectively designated by the acronym SH and CSPR. It is worth to emphasize that the comments are based on the simple observation. For a more detailed analysis, it would be necessary the application of a classifier in order to investigate the accuracy of classification regarding both groups of selected variables by the MIFS-U.

Regarding the ECHOCARDIOGRAM database (Table 2), the selection made by the MIFS-U using the two methods leads to two similar sets of selected variables. Three among the first five variables selected by the algorithm are exactly the same. It is noteworthy that the possibility exists that the variables 2 and 3 selected using the SH method have contribution for the output similar to the one of the variables 5 and 9 selected using the CSPR method. In practical terms, it would mean that in principle the permutation of these subsets in the set of selected variables would have little influence on the result (that is, the classification) that must be ascertained by application of a classifier.

Table 2 – Comparative result of the selection by the MIFS-U – ECHOCARDIOGRAM.

ECHOCARDIOGRAM Database				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1st	4	1.0000	4	1.0000
2nd	1	0.7424	1	0.8705
3rd	2	0.0275	10	0.2123
4th	3	0.0258	5	0.1324
5th	10	0.2963	9	0.1246

Table 3 – Comparative result of the selection by the MIFS-U – TELESCOPE.

Database TELESCOPE				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1st	1	1.0000	9	1.0000
2nd	9	1.0000	2	0.1874
3rd	8	0.9963	1	0.1441
4th	10	1.0000	7	0.1029
5th	6	0.9963	3	0.0661

Regarding the TELESCOPE database (Table 3), the selection made by the MIFS-U using the two methods is again practically the same in relation to the first three variables selected by the algorithm. Equally, the possibility exists that the remaining variables in the two sets of selected variables have similar contibution for the output, but this must be checked by application of a classifier. It is noteworthy that, using the SH method, the mutual information of each selected variable with respect to the output is practically the same, which does not happen using the CSPR method, reflecting a more discriminating power.

Regarding the WINE database (Table 4), the selection made by MIFS-U using the two methods is again quite similar, as it happened with the ECHOCARDIOGRAM database. Thus the same considerations hold. Obviously, the application of a classifier would provide a clear vision of the contribution of the remaining variables, in relation to both methods. It is noteworthy once again that, using the SH method, the mutual information of each selected variable with respect to the output is practically the same, which does not happen using the CSPR method, reflecting a more discriminating power.

Table 4 – Comparative result of the selection by the MIFS-U – WINE.

Database WINE				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1^{st}	1	1.0000	13	1.0000
2^{nd}	2	0.9942	1	0.9945
3^{rd}	5	0.9942	10	0.7680
4^{th}	3	0.9942	2	0.3011
5^{th}	11	0.9942	5	0.4337

Table 5 – Comparative result of the selection by the MIFS-U – DERMATOLOGY.

Database DERMATOLOGY				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1^{st}	16	1.0000	16	1.0000
2^{nd}	23	0.9820	18	0.9703
3^{rd}	28	0.4910	17	0.8514
4^{th}	27	0.0048	20	0.5053
5^{th}	22	0.0048	7	0.6242

Regarding the DERMATOLOGY database (Table 5), the variable 16 was the first variable selected by the MIFS-U using both methods. However the variable 23, second variable selected by the algorithm using the SH method, is one of the last ones classified using the CSPR method. The inverse happens with the variable 18. Although the mutual information of each one of these variables with respect to the output is significant, the possibility exists that these variables are redundant, what must be examined in more detail. For both methods, the mutual information between them is considerable, that is, 0,852(SH) and 0,771(CSPR).

Regarding the BREAST CANCER database (Table 6), the variables 23 and 28, although in inverted order, were the first variables selected by the MIFS-U using both methods. It can be still observed that the other three variables, in relation to both methods, except the variable 14 in the SH case, have very low mutual information with the output, indicating probably a particular contribution of these variables.

Regarding the HEART DISEASE database (Table 7), the selection made by the MIFS-U using the two methods is once again practically the same. It is possible that the variables 5 and 7 selected using the SH and CSPR methods, respectively, have similar contribution for the output, observing that, in the set of the first five variables selected by the algorithm, only the variables 5 and 7 determine the difference between the methods.

Table 6 – Comparative Result of the Selection by the MIFS-U – BREAST CANCER.

Database BREAST CANCER				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1st	23	1.0000	28	1.0000
2nd	28	0.9520	23	0.8659
3rd	14	0.6463	20	0.0843
4th	17	0.2722	12	0.0268
5th	2	0.2969	29	0.1648

Table 7 – Comparative Result of the Selection by the MIFS-U – HEART DISEASE.

Database HEART DISEASE				
Order	SH Method		CSPR Method	
	Var.	MI with Output	Var.	MI with Output
1st	9	1.0000	2	1.0000
2nd	8	0.9864	8	0.5806
3rd	2	0.9676	9	0.4194
4th	7	0.7022	1	0.3871
5th	5	0.6823	7	0.3065

6 FINAL REMARKS

Variable selection is a fundamental problem in several areas of knowledge. All the variables may be important within a given context, but for a particular concept, only a small subset of variables is usually relevant. Besides, variable selection increases the intelligibility of a model, while reducing the dimensionality and the need for storage space. Several experimental studies have shown that irrelevant and redundant variables can drastically reduce the predictive accuracy of models built from data. In this paper, the Mutual Information Variable Selector under Uniform Information Distribution (MIFS-U) was approached. This algorithm, as was shown, involves the computation of entropy and mutual information regarding discrete and continuous variables. In the first case, the computation is straightforward, but for continuous variables, there are inevitable integrals in all the definitions of entropy and mutual information, which are the major difficulty after the density estimation. Therefore the density estimation and measures of entropy and mutual information should be chosen appropriately so that the corresponding integrals can be simplified. It was shown how the Rényi quadratic entropy and the Cauchy-Schwartz quadratic mutual information, instead of the Shannon entropy and Shannon mutual information, can be combined with the Gaussian kernel function to estimate densities, resulting in an effective and general method for computing entropy and mutual information, without requiring any hypothesis about the unknown density – in almost all real world problems, the only information available is contained in the data collected. It should be always kept in mind that the process of variables selection must be as accurate as possible, but without losing its simplicity. In practice, simplicity becomes a paramount consideration. If such process involves complex techniques, it ends up becoming a problem in itself, rather than being a facilitator for a later stage of classification, through, for example, learning of an Artificial Neural Network (ANN).

Experiments were conducted, comparing the Cauchy-Schwartz / Parzen-Rosenblatt method (CSPR), presented in this paper, with the Shannon/Histogram method (SH), widely used, based on the Shannon entropy definition and that uses the discretization of continuous variables as a step of pre-processing of the data. The results, focusing on the set of the first five selected variables, were similar. As the comparison was purely speculative, a more careful analysis must be realized by applying a classifier (or more than one), so that the methods can be compared through the effective performance of the sets of selected variables by the MIFS-U algorithm. Besides, it is strongly recommended the participation of a professional in the field of knowledge concerning the databases covered in this paper, as it would certainly allow a better evaluation of the methods. Lastly, the CSPR method works directly with the data, providing, theoretically, greater accuracy. On the other hand, the SH method – that uses the discretization, which in principle could mask some relevant "information" from the data – is simpler, which explains its widespread use.

REFERENCES

[1] Agrawal R, Imielinski T & Swami A. 1993. Database minimng: A performance perspective. *IEEE Trans. Knowledge Data Eng.*, **5**: December 1993.

[2] Battiti R. 1994. Using mutual information for selecting features in supervised Neural net learning. *IEEE Trans. Neural Networks*, **5**: 537–550.

[3] Breiman L et al. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[4] Cavalcante CC. 2001. Predição Neural e Estimação de Função Densidade de Probabilidades Aplicadas à Equalização Cega. Dissertação de Mestrado, DEE/UFC.

[5] Cover TM & Thomas JA. 2006. *Elements of information theory*. 2nd ed., Jonh Wiley & Sons, Inc., Hoboken, New Jersey.

[6] Duda RO & Hart PE. 1973. *Pattern Classification and Scene Analysis*. Jonh Wiley & Sons, Inc., New York.

[7] Hartley RV. 1928 Transmission of Information. *Bell System Technical Journal*, **7**: 535–563, July 1928.

[8] Joliffe IT. 1986. *Principal Component Analysis*. Springer-Verlag, New York.

[9] Kullback S. 1968. *Information Theory and Statistics*. Dover Publications, Inc., New York.

[10] Kwak N & Choi C. 2002. Input Feature Selection for Classification Problems. *IEEE Trans. Neural Networks*, **13**(1): 143–159.

[11] MACRINI JLR. 2004. Estimação do Risco de Recidiva em Crianças Portadoras de Leucemia Linfoblástica Aguda Usando Redes Neurais. Tese de Doutorado, DEE/PUC-Rio.

[12] NEEMUCHWALA HF. 2005. Entropic Graphs for Image Registration. PhD thesis, University of Michigan.

[13] PRINCIPE JC ET AL. 2000. Learning from examples with information theoretic criteria. *Journal of VLSI Signal Proc. Systems*, **26**(1/2): 61–77, August 2000.

[14] PRINCIPE JC, FISHER III JW & XU DX. (1998). *Information-Theoretic Learning*. University of Florida, Gainesville.

[15] RODRIGUES TB. 2006. Seleção de Variáveis e Classificação de Padrões Utilizando Redes Neurais com Aplicação no Diagnóstico de Doença Cardíaca. Dissertação de Mestrado, DEE/PUC-Rio.

[16] SCOTT DW. 1992. *Multivariate Density Estimation.* Jonh Wiley & Sons, Inc., New York.

[17] SHANNON CE & WEAVER W. 1949 *The Mathematical Theory of Communication*. Univ. Illinois Press, Urbana, IL.

[18] SILVEIRA GB. 1992. *Estimação de Densidades e de Funções de Regressão*. UFRJ, Rio de Janeiro.

[19] SILVERMAN BW. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[20] VIOLA P & WELLS III WM. 1997. Alignment by Maximization of Mutual Information. *International Journal of Computer Vision*, **24**(2): 137–154.

[21] WAND MP & JONES MC. 1995. *Kernel Smoothing*. Chapman and Hall, London.

[22] XU D. 1999. Energy, Entropy and Information Potential for Neural Computation. PhD thesis, University of Florida.

[23] XU D ET AL. 1998. A novel measure for independent component analysis (ica). *IEEE International Conference on Acoustics, Speech and Signal Processing*, **2**: 1145–1148.