Application of Rényi Entropy and Mutual Information of Cauchy-Schwartz in Selecting Variables

Leonardo Macrini[1], Leonardo Gonçalves[2]

[1] Departamento de Ciências Econômicas e Exatas Universidade Federal Rural do Rio de Janeiro. Três Rios, RJ, Brasil E-mail: macrini@centroin.com.br

[2] Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro, RJ, Brasil.

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.

1 Introduction

Variable selection has a fundamental importance in classification systems, such as Neural Networks [1], [2], [3]. In this paper, the Mutual Information Variable Selector under Uniform Information Distribution (MIFS-U) is focused [4]. 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 [5]. When the variables involved are discrete, the computation of entropy and mutual information, based on the Shannon definition, is simple 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 [6], 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) [7], [8]. 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 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 [2] 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 ← "initial set of n variables", S ← "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 ← F \{$\phi_i$}, S ← {$\phi_i$}.
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 ← F \ {$\phi_i$}, S ←{$\phi_i$}.
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).

2.1 The MIFS-U Variable Selector

The ideal algorithm [2] tries to maximize $I(D;\phi_i,\phi_s)$ (area II, III and IV in Figure 1) and, according to [4], this can be rewritten as

$$I(D;\phi_i,\phi_s) = I(D;\phi_s) + I(D;\phi_i \mid \phi_s). \tag{1}$$

Where $I(D;\phi_i \mid \phi_s)$ 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 1, 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 $I(D;\phi_i \mid \phi_s)$ (area III). However, calculating $I(D;\phi_i \mid \phi_s)$ requires as much work as calculating *I*(*D*; ϕ_i; ϕ_s). So $I(D;\phi_i \mid \phi_s)$ 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

$$I(D;\phi_i \mid \phi_s) = I(D;\phi_i) - \{I(\phi_i;\phi_s) - I(\phi_i;\phi_s \mid D)\} \tag{2}$$

Where $I(\phi_i;\phi_s)$ corresponds to arera I and IV, and $I(\phi_i;\phi_s \mid D)$ corresponds to area I. So the term $I(\phi_i;\phi_s) - I(\phi_i;\phi_s \mid D)$ corresponds to area IV. The term $I(\phi_i;\phi_s \mid D)$ 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):

$$\frac{H(\phi_s \mid D)}{H(\phi_s)} = \frac{I(\phi_i;\phi_s \mid D)}{I(\phi_i;\phi_s)} \tag{3}$$

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

$$I(\phi_i;\phi_S \mid D) = \frac{H(\phi_s \mid D)}{H(\phi_s)}\, I(\phi_i;\phi_s) \tag{4}$$

Using the equation above and Eq. (2), the following is obtained:

$$I(D;\phi_i \mid \phi_s) = I(D;\phi_i) - \frac{I(D;\phi_s)}{H(\phi_s)}\, I(\phi_i;\phi_s). \tag{5}$$

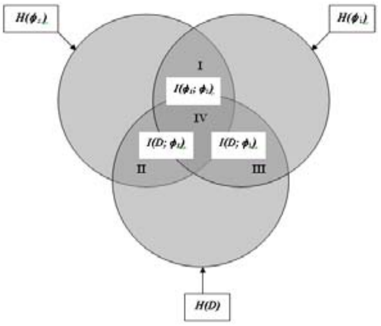

Fig. 1 – The relation between input variables and output classes

Assuming that each region in Figure 1 corresponds to its corresponding information, the condition presented in Eq. (3) is hard to satisfied when information is concentrated on one of the following regions: $H(\phi_s \mid \phi_i; D)$, $I(\phi_s; \phi_i \mid D)$, $I(D; \phi_s \mid \phi_i)$ or $I(D; \phi_s; \phi_i)$. It is more likely that condition (3) hods when information is distributed uniformly throughout the region of $H(\phi_s)$ in Figure 1. 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 $I(D;\phi_i) - \beta \sum_{\phi_s \in S} \big(I(D;\phi_s) / H(\phi_s)\big) I(\phi_i;\phi_s)$ and set $F \leftarrow F \setminus \{\phi_i\}$, $S \leftarrow \{\phi_i\}$.

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 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 [6]. 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 [4] point out that the MIFS-U algorithm can be applied to large problems without excessive computational efforts.

3 Estimation Methods of Entropy and Mutual Information

3.1 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 [10]. 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 [11].

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 apllied.

3.2 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 [14], integrated with the Parzen Window estimator.

The Kullback-Leibler divergence [14], 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 [8]. Xu et al. [12] presented a method that combines the Cauchy-Schwartz Divergence with Parzen Windowing for estimating the mutual information directly from the data.

Further details of the calculations required in this process can be found in the work of Gonçalves & Macrini [16].

4 Experiments

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.

Table 1 – Databases

Databases	**n**	**Number of Variables**	
		Discrete	**Continuous**
ECHOCARDIOGRAM			

Echocardiogram Data	61	3	8
BREAST CANCER ***Wisconsin Diagnostic Breast Cancer***	569	0	30

4.2 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.

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

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 3 – Comparative Result of the Selection by the MIFS-U - BREAST CANCER

Database BREAST CANCER		
	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

Regarding the BREAST CANCER database (Table 3), 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.

5. 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 Joliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

4 Kwak, N.; Choi, C. (2002). Input Feature Selection for Classification Problems. *IEEE Trans. Neural Networks*, **13**(1), 143-159.

5 Cover, T. M.; Thomas, J. A. (2006). *Elements of information theory*. 2nd ed., Jonh Wiley & Sons, Inc., Hoboken, New Jersey.

6 Breiman, L. et al. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

7 Principe, J. C. et al. (2000). Learning from examples with information theoretic criteria. *Journal of VLSI Signal Proc. Systems*, **26**(1/2), 61-77, August 2000.

8 Principe, J. C.; Fisher III, J. W.; Xu, D. X. (1998). *Information-Theoretic Learning*. University of Florida, Gainesville.

9 Hosmer & Lemeshow (1989). Applied Logistic Regression. John Wiley & Sons, New York.

10 Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

11 Scott, D. W. (1992). *Multivariate Density Estimation.* Jonh Wiley & Sons, Inc., New York.

12 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.

13 Wand, M. P.; Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

14 Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications, Inc., New York.

15 Shannon, C. E.; Weaver, W. (1949) *The Mathematical Theory of Communication*. Univ. Illinois Press, Urbana, IL.

16 Gonçalves, L.B.; Macrini, L. (2011). Rényi Entropy and Cauchy-Schwartz Mutual Information Applied to MIFS-U Variable selection Algorithm: A Comparative Study. Pesquisa Operacional. Brazilian Operations Research Society, 31(3), 499-519.