On the minimum probability of classification error through effective cardinality comparison

Jugurta Montalvão, Jânio Canuto, and Elyson Carvalho

Abstract—**This work proposes a method for estimating a lower (Bayesian) bound for classification error rate in two-class problems. This lower bound is typically inferred through specific classifier structures. By contrast, the proposed approach is based on "collision" (quadratic) entropy estimator, deployed in the very pragmatic form of simple coincidence counters. To properly introduce this new approach, we first discuss the concept of sets' effective cardinality in view of basic concepts of probability and set theory. The usefulness as well as some limitations of the analysis based on effective cardinality are exemplified throughout the text.**

Index Terms—**Entropy through coincidence, Multivariate statistics, Collision entropy, Effective cardinality, Classification error probability.**

I. Introduction

A fundamental problem in pattern recognition is the optimization of classifiers. Whenever the criterion is the minimization of probability of misclassification, it is well known that the optimum performance is reached by the Bayesian classifier [1]. However, in many practical situation, an experimenter may only have few data samples, possibly multivariate and/or heterogeneous, so that the Bayesian classifier becomes an unattainable goal, for it depends on the perfect knowledge of classes probability distributions and respective *a priori* probabilities.

Consequently, the minimum error that the idealized Bayesian classifier can provide is a lower bound that cannot be perfectly known in such practical cases. Approximated values for this lower error bound are generally obtained through indirect approaches, such as classifiers with enough degrees of freedom to empirically find suitable, possibly nonlinear, classification boundaries. Good examples of these practical classifiers are artificial neural networks and support vector machines. These approaches allow for the adjustment of classification boundaries without explicit class distribution estimation. A Bayesian classifier can also be implemented by explicit estimation of the probability distribution of each class, followed by the application of the maximum *a posteriori* probability decision rule. In both cases, however, the minimum classification error is estimated as a consequence of a classification structure applied to the available data.

By contrast, a similar problem in telecommunications historically received a more direct approach, namely, the problem of minimizing transmission error. To tackle this problem, C. Shannon [2] discussed the probability of signals being wrongly associated to symbols – symbols used in a digital communication scheme –, and successfully established relationships between quantities of symbols in sets (cardinalities) and error rates. His work was not the first to address the problem in terms of cardinalities, for as early as in 1928 (thus before the seminal papers published by Shannon), Ralph V. L. Hartley was already concerned with a practical measure of information given by the logarithm of the number of possible symbol sequences [3], or the cardinality of a set of messages, which later on was directly linked to the concept of entropy, in information theory. In 1951, a short but important note by S. Kullback and R. A. Leibler [4] generalized the Shannon entropy as a measure of divergence between statistical populations. The so-called Kullback-Leibler divergence was proposed as a tool to measure how difficult it is to discriminate between two competing distributions, in terms of measure of information. S. Kullback and R. A. Leibler exemplified the usefulness of their divergence through its application on testing a given hypothesis, which turns out to be a matter of measuring probability of errors (*i.e.* false acceptance and false rejection of a hypothesis). This divergence was further generalized by the Rényi's family of divergences [5].

Finding a noisy channel capacity or discriminating between two competing distributions are practically motivated problems closely related to finding the minimum probability of classification error. But it is noteworthy that the pragmatic motivation of works such as those by Shannon or by Hartley migrated, through the years, to sophisticated theoretical developments, not always aimed at practical applications, as pointed out in [6]. Nevertheless, nowadays, it is widely agreed that entropy based tools are useful in most domains where information can be symbolically and/or numerically handled, ranging from Econometrics [7] to Biology [8].

In this work, we propose an approach to estimate the minimum probability of classification error through cardinality of sets and simple probability concepts, in the manner of early works on information theory. Moreover, in order to keep our approach as pragmatic as possible, we also follow the ideas proposed by M. O. Hill [8], in which cardinalities are regarded as measures of the degree of specie polydominance, in Biology. We adhere to their standing point as much as possible, thus replacing the less intuitive concept of entropy with the mirror concept we chose to call *effective cardinality* (defined in Section II).

We take advantage of the fact that effective cardinalities are easier to handle than entropy, and we explain, in Section II, the principles of the proposed method as plainly as possible in terms of cardinality comparisons. The method itself is

J. Montalvão, Jânio Canuto, and Elyson Carvalho are with the Federal University of Sergipe (UFS), São Cristóvão, Brazil e-mail: jmontalvao(at)ufs.br.

Digital Object Identifier: 10.14209/jcis.2016.26

presented in Section III and, since it relies upon a proper definition of the probabilistic event *coincidence* (or *collision*), this sensitive point is separately discussed in Subsection III-A, before the conclusions presented in Section IV.

II. Classification error, entropy and effective cardinality

In this Section, the principles upon which the method proposed in Section III is based are presented through examples. We reckon that it is a rather unusual paper structure, but we believe that it is more suitable to the reader, in terms of presentation of practical aspects behind some abstract concepts.

Accordingly, the first example just shows how set cardinality and classification error probability are related for random variables with uniform distributions. Step by step, further examples are added to gradually explain how quadratic entropy can be used as a tool to adapt this principle even to nonuniform random variables, through the concept of effective cardinality.

As for the first example, we consider random variables defined in terms of sets whose elements are drawn with equal probabilities. In this case, there is a straightforward relationship between set cardinality and classification error probability.

Example 1: Two uniform random variables, X and Y, take values from sets $\mathcal{X} = \{1,\ 2,\ 3,\ 4\}$ and $\mathcal{Y} = \{2,\ 3,\ 4,\ 5,\ 6\}$, respectively, whose cardinalities are $C_X = 4$ and $C_Y = 5$. If an instance, z, is observed, but the observer does not know whether it is an instance of X or Y (The observer only knows that z is equally likely to be an instance of X or Y), three scenarios are possible, namely:

(a) $z \in \{1\}$, and it is decided that z is an instance of X.
(b) $z \in \{5,\ 6\}$, and it is decided that z is an instance of Y.
(c) $z \in \{2,\ 3,\ 4\}$ and it is decided that z is an instance of X (because $\frac{1}{C_X} > \frac{1}{C_Y}$, see explanation below).

Clearly, only in the later scenario an error may occur, and to minimize the error probability, decision (c) corresponds to the Bayesian criterion according to which a decision might be made in favour of X or Y by comparing the probabilities $P_X \Pr(z|X)$ and $P_Y \Pr(z|Y)$, where P_X and P_Y are *a priori* probabilities of z being an instance of X or Y, respectively. Given that $P_X = P_Y$, $\Pr(z|X) = 1/C_X$ and $\Pr(z|Y) = 1/C_Y$, the observer should systematically decide that any z from the intersection of the two sets is an instance of X, because $1/C_X > 1/C_Y$.

In this example, the minimum probability of error is $\Pr(error) = P_Y \Pr(z \in \mathcal{X} \cap \mathcal{Y}|Y) = P_Y \frac{C_I}{C_Y} = 0.3$, where $C_I = 3$ stands for the cardinality of the intersection between $\mathcal{X}$ and $\mathcal{Y}$, and the cardinality comparison C_I/C_Y is the probability of an element of the intersection be drawn, given that z is an instance of Y (thus a decision error, since the observer should systematically decide in favour of X).

By generalizing this approach for uniformly distributed random variable, the minimum error probability is given by

$$\Pr(error) = C_I \min\left(\frac{P_X}{C_X}, \frac{P_Y}{C_Y}\right). \quad (1)$$

Example 2: X and Y are the same uniform random variables defined in Example 1, but the respective *a priori* probabilities are changed to $P_X = 1/3$ and $P_Y = 2/3$. In this case, for scenario (c), it should be decided that z is an instance of Y (because $\frac{P_Y}{C_Y} > \frac{P_X}{C_X}$), and the minimum probability of error is now $\Pr(error) = C_I \frac{P_X}{C_X} = 0.25$

The approach illustrated through examples 1 and 2 are based on very basic concepts of probability and set theory, and the simplicity of the proposed formula to compute error probability through cardinalities comparison (Equation 1) comes from the fact that we are handling finite sets of equiprobable elements.

To expand this straightforward approach to other kinds of random variables, including those with continuous cumulative distribution function, we use the concept of *effective cardinality*, which is itself a proxy to the concept of entropy, as explained in [8], with different terminology. Accordingly, we adhere to the point of view presented in [8], through which we further conjecture that a key point to explain – and to use – the entropy of a random variable X as plainly as possible, mainly for those with mathematical background under development, is to replace it with the notion of *effective number of elements* of the event space $\mathcal{X}$. This number is not expected to be the actual cardinality of $\mathcal{X}$, but rather an *effective cardinality* of a (equivalent) set with equiprobable events. As a consequence, the *effective cardinality* may even be a fractional number. In [8], from the perspective of applications in Ecology, this quantity was referred to as the "effective number of species present in a sample", or "diversity number of order a", given by

$$C^{(a)} = \left(\sum_{i=1}^{K} w_i p(i)^{(a-1)}\right)^{\frac{1}{1-a}}, \quad (2)$$

where K is the total number of detected species in the sample and $w_i = p(i)$ is the proportion of specie i (thus $\sum_{i=1}^{K} p(i) = 1$ and $p(i) \geq 0,\ \forall i$). This formulation by M. O. Hill allows a thoughtful understanding of $C^{(a)}$ as a weighted generalized mean, or power mean with exponent $(a-1)$, thus an average quantity. Moreover, by replacing the proportion of species, $p(i)$, with the probability of the *i-th* event to occur, Equation 2 also provides an interesting standing point according to which the entropy generalization in [5], $H^{(a)}$, is just the logarithm of the average quantity $C^{(a)}$, including the *Shannon entropy*, for $a = 1$, probably the most used entropy definition in information theory.

On the other hand, *collision entropy* ($a = 2$) is advantageous in practical experiments because it gathers all detected coincidences (or collisions) in a single counter, as opposed to histogram based approaches. As a matter of fact, entropy estimators are always based on some kind of coincidence detection, explicitly or not. For instance, simple *plug-in* approaches for entropy estimation typically use histograms as estimators of probability mass functions (PMF), and then use the resulting estimates as actual PMFs in entropy formulas. We highlight that to build up histograms it is necessary to define a set of reference symbols (histogram bins), and to count coincidences between observed instances and elements of this reference set. Therefore, each bin in a histogram can be regarded as a counter for a specific kind of coincidence. By contrast, in *collision* based approaches, all detected coincidences between pairs of observed instances are gathered into a single counter. As pointed out by researchers such as A. Bialas and W. Czyz [9], this is experimentally attractive because the statistical error of $C^{(2)}$ estimates drops very fast (inversely proportional to the number of available instances). This fast statistical convergence is convenient for practical purposes where datasets are limited in size, and it is directly related to what I. Nemenman [10] called the Ma square-root regime, as a reference to the findings of S.-K. Ma [11].

Therefore, due to the pragmatic flavour of this work, we narrow our attention to the quadratic entropy, although we also keep an eye on the numerical results provided by the effective cardinality of order one, $C^{(1)}$, which is related to the Shannon definition of entropy. Accordingly, we define:

> **Definition:** $C_X^{(a)}$ is the effective cardinality of order a associated to a random variable X.

For a discrete random variable, the effective cardinality is given by Equation 2. In the specific case of $a = 2$, $C_X^{(2)}$ is the number of elements in the support set of an idealized uniform random variable W, so that two independent instances of W have the same probability of collision – or coincidence – as two independent instances of X.

Given this definition, one may wonder if the simple formulae presented in Equation 1 would hold for idealized uniform sets associated to nonuniform random variables X and Y. We address this question in the next examples. Before, however, it should be noticed that, unlike the effective cardinalities of X and Y, the effective cardinality of the intersection is not directly available. Fortunately, as illustrated in Example 3, $C_I^{(2)}$ can be estimated through the counting of coincidences between independent instances of X and Y.

Example 3: The total number of pairs where one element is taken from $\mathcal{X} = \{1,\ 2,\ 3,\ 4\}$, and the other is taken from $\mathcal{Y} = \{2,\ 3,\ 4,\ 5,\ 6\}$ is $C_X C_Y = 20$. Amongst these 20 pairs, there are $C_I = 3$ formed by coincident elements, clearly corresponding to the cardinality of the intersection $\mathcal{X} \cap \mathcal{Y}$. The probability of randomly finding one of these 3 coincident pairs amongst the total of 20 is

$$\Pr(X = Y) = \frac{C_I}{C_X C_Y} = \frac{3}{20}. \tag{3}$$

To keep the analogy with the probabilities of coincidence between instances of X or Y, separately, which are $1/C_X$ and $1/C_Y$, respectively, we define a new cardinality, C_{XY}, such as

$$\Pr(X = Y) = 1/C_{XY}, \tag{4}$$

where C_{XY} may be regarded as a cross-cardinality. In this illustration with equally probable elements, C_I is at hand, and $C_{XY} = 20/3$ is a rather unnecessary and almost meaningless quantity. But, in cases where random variables are nonuniform, C_{XY} becomes a useful measure that can be directly estimated trough cross-coincidence counting, whereas C_I is not directly available. Indeed, from 3 and 4 we may obtain C_I as a function of all the cardinalities directly estimable through coincidence counting, namely:

$$C_I = C_X C_Y / C_{XY}.$$

This definition is generalized to effective cardinalities of order a as

$$C_I^{(a)} = \frac{C_X^{(a)} C_Y^{(a)}}{C_{XY}^{(a)}}. \tag{5}$$

where we also define cross-cardinality of order a as a generalization of Equation 2, as

$$C_{XY}^{(a)} = \left(\sum_{i \in \mathcal{X} \cup \mathcal{Y}} p_X(i) p_Y(i)^{(a-1)} \right)^{\frac{1}{1-a}}, \tag{6}$$

for discrete variables, where p_X and p_Y stand for the PMF of X and Y, respectively, and

$$C_{XY}^{(a)} = \left(\int_{z \in \mathcal{X} \cup \mathcal{Y}} f_X(z) f_Y(z)^{(a-1)} \right)^{\frac{1}{1-a}} dz, \tag{7}$$

for continuous variables, where f_X and f_Y stand for the probability density functions (PDF) of X and Y, respectively.

To illustrate how fair is the minimum error probability estimated trough effective cardinality analysis on nonuniform random variables, Example 4 addresses a case that can be regarded as a model for two unfair dices being independently thrown. Please note that the former examples can be regarded as specific cases of Example 4.

Example 4: Two random variables, X and Y, take values from sets $\mathcal{X} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ and $\mathcal{Y} \equiv \mathcal{X}$, with probabilities $p_X(z) = \Pr(X = z)$ and $p_Y(z) = \Pr(Y = z)$ respectively. An instance, z, is observed, and the observer only knows p_X, p_Y, as well as that z is likely to be an instance of X or Y with probabilities $P_X = P_Y = 0.5$. For this example, 100,000 instances of distributions p_X and p_Y were randomly created. A validated pseudorandom number generator was used to generate values for $p_X(z)$ and $p_Y(z)$ between 0 and 1, with uniform density. These values were further normalized for p_X and p_Y to be a valid PMF. The minimum error probability, $\Pr(error)$ (according to the Bayesian criterion) was compared to $Pe^{(2)} = C_I^{(2)} \min\left(\frac{P_X}{C_X^{(2)}}, \frac{P_Y}{C_Y^{(2)}}\right)$, thus yielding, in average, $\Pr(error) = 0.338$ and $Pe^{(2)} = 0.355$.

For sake of comparison with the cardinality associated to the Shannon entropy, we also computed $Pe^{(1)} = C_I^{(1)} \min\left(\frac{P_X}{C_X^{(1)}}, \frac{P_Y}{C_Y^{(1)}}\right)$, which yielded, in average, $Pe^{(1)} = 0.316$. Therefore, we observed that $Pe^{(1)}$ is biased by about -0.022 , whereas $Pe^{(2)}$ bias is about 0.017. Besides, the standard deviations of errors $(Pe^{(1)} - \Pr(error))$ and $(Pe^{(2)} - \Pr(error))$ are, respectively, 0.058 and 0.023.

In this case with random variables taking values in a set of six elements, both $Pe^{(2)}$ and $Pe^{(1)}$ are biased estimators of $\Pr(error)$, with a bias of about 5% of the actual error probability for $Pe^{(2)}$. To further explore the limits of these two rough probability estimators, we extended Example 4 to X and Y taking values from the much bigger set: $\{1, 2, 3, \ldots, 1000\}$, which yielded biases for $Pe^{(1)}$ and $Pe^{(2)}$ of about -0.03 and 0.04, with standard deviations of about 0.006 and 0.002, respectively.

Examples from 1 to 3 are aimed at building an intuitive perception that the minimum error probability can be roughly estimated through effective cardinality analysis, even for nonuniform random variables, and the experimental results from Example 4 corroborate this perception. Moreover, the relatively small deviations obtained for $Pe^{(2)}$ is the most welcome, because a quadratic entropy/cardinality estimator is known to provide meaningful results even for small sets of random variable instances [9], as compared to most Shannon entropy estimators.

Up to this point, we already know that, for classes modelled as discrete random variables, the minimum classification error probability can be roughly estimated through effective cardinality analysis. However, a more difficult to handle classification problem may include continuous random variables. Fortunately, the concept of effective cardinality still holds for continuous variables, and to study the approach under continuous variables, we do two brief experiments, namely:

1) First we consider X and Y as two continuous random variables with uniform probability density functions, f_X and f_Y, with unit range ($f_X(z) = 1$ if $|z - \mu_X| < 0.5$ and $f_Y(z) = 1$ if $|z - \mu_Y| < 0.5$), centered at μ_X and μ_Y, respectively. The *a priori* probabilities are arbitrarily set to $P_X = 0.4$ and $P_Y = 0.6$, just to improve graphic visualization, as in Fig. 1.
2) Secondly we consider X and Y as two Gaussian random variables, whose probability density functions, f_X and f_Y, are two unit variance Gaussians with means μ_X and μ_Y, respectively. Their *a priori* probabilities are the same (i.e. $P_X = P_Y = 0.5$), as represented in Fig. 2.

In both experiments, the actual minimum classification error (for the optimum Bayesian classifier) is also computed while the gap $\mu_Y - \mu_X$ between the two classes is gradually increased. Error estimates – through effective cardinality analysis – are obtained according to:

- $\hat{Pe}^{(1)} = C_I^{(1)} \min\left(\frac{P_Y}{C_Y^{(1)}}, \frac{P_X}{C_X^{(1)}}\right)$;
- $\hat{Pe}^{(2)} = C_I^{(2)} \min\left(\frac{P_Y}{C_Y^{(2)}}, \frac{P_X}{C_X^{(2)}}\right)$,

where effective cardinalities are obtained according to Equation 7, with notation simplified as

$$C_X^{(1)} = C_{XX}^{(1)}; \tag{8}$$

$$C_Y^{(1)} = C_{YY}^{(1)}; \tag{9}$$

$$C_X^{(2)} = C_{XX}^{(2)}; \tag{10}$$

$$C_Y^{(2)} = C_{YY}^{(2)}; \tag{11}$$

and $C_I^{(a)} = C_X^{(a)} C_Y^{(a)} / C_{XY}^{(a)}$.

Through this experiment, we observe that the estimator based on quadratic effective cardinality gives a precise estimate of $\Pr(error)$ for classes with uniform distribution, whereas the estimator based on order one cardinalities (related to the Shannon entropy) clearly fails[1]. These results strongly corroborates the intuition that the quadratic entropy, being related to *collision* detection – which turns out to be a matter of fundamental comparison between instances of random variables – would also be more closely related to classification error. Likewise, in the second experiment, whose results are presented in Fig. 2, the estimate of order two seems to be a more suitable choice for pattern classification purposes, giving better probability estimates for values of $\Pr(error)$ below 10%.

However, considering Equations 6 and 7, error estimation through effective cardinality, at a first glance, is just a rough estimate whose computational burden is equivalent to that of the direct estimation of $\Pr(error)$. Therefore, if distributions

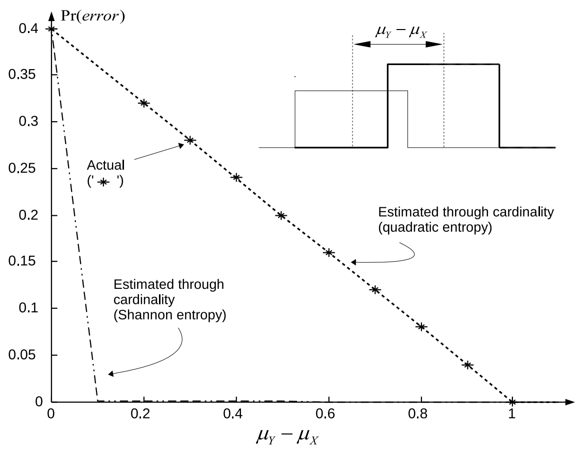


Fig. 1. Comparison of the actual error probability to its estimates through cardinalities for $a = 1$ (corresponding to the Shannon entropy) and $a = 2$ (corresponding to the quadratic entropy) – Two classes with uniform distributions.

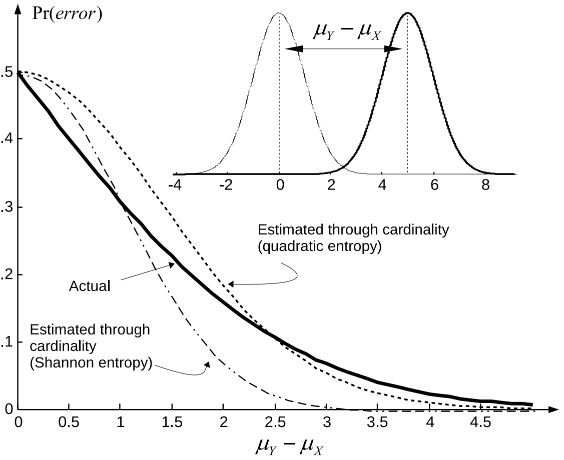


Fig. 2. Comparison of the actual error probability to its estimates through cardinalities for $a = 1$ (corresponding to the Shannon entropy) and $a = 2$ (corresponding to the quadratic entropy) – Two classes with Gaussian distributions.

[1] $C_{XY}^{(1)}$ cannot even be properly computed if $f_Y(z) = 0$ for any z such that $f_X(z) \neq 0$. To avoid singularity, $f_Y(z)$ was replaced with $f_Y(z) + \varepsilon$, with $\varepsilon \approx 0$.

are known, as well as the classes *a priori* probabilities, P_X and P_Y, either $\Pr(error)$ or $Pe^{(2)}$ are obtained through sum/integration of functions over the support sets of X or Y. For instance, if both X and Y are continuous random variables, the computation of $\Pr(error)$ is given by integrals of $P_X f_X(z)$ or $P_Y f_Y(z)$, over regions R_Y and R_X, respectively, where R_X (respectively R_Y) is a region where $P_X f_X(z) \geq P_Y f_Y(z)$ (resp. $P_X f_X(z) < P_Y f_Y(z)$). Thus, in terms of computational burden, there would not be an evident reason for using the proposed rough estimator $Pe^{(2)}$. But a known caveat that may hinder the computation of $\Pr(error)$ is that R_Y and R_X may be formed by imbricated disjoint subregions. For instance, the one-dimensional random variables whose probability density functions are represented in Fig. 3, under equal *a priori* probabilities (for simplicity), form optimum classification regions which are disjoint. For this one-dimensional problem, the minimum classification error probability is still easily computable, yielding $\Pr(error) = 0.092$, but it gives an idea of how difficult it would be if X and Y were multivariate, thus defined in high-dimensional spaces and forming irregular and disjoint regions R_X or R_Y, difficult to integrate over, mainly because of the issue of defining the irregular integration boundaries.

By contrast, effective cardinalities are always computed over the entire support set of X and Y, which can be much simpler in many practical cases, and yet, it may give a useful estimate of $\Pr(error)$. In the example corresponding to the distributions in Figure 3, $Pe^{(2)} = 0.089$ whereas $\Pr(error) = 0.092$.

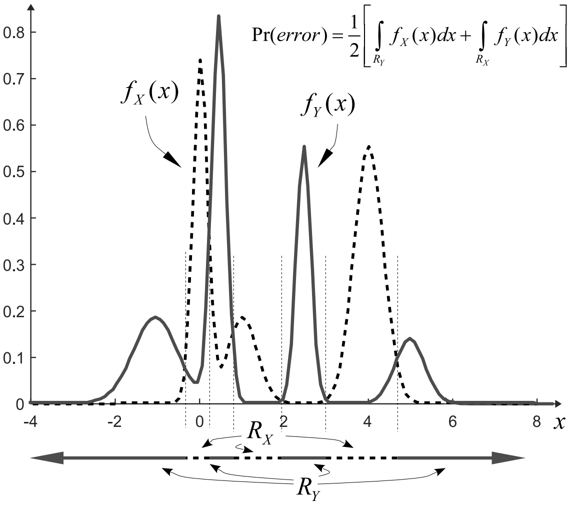


Fig. 3. Bayesian optimum classifier illustrated for a unidimensional two-class problem.

It should be noticed that the computation of $\Pr(error)$ is prone to strong deviations in practical situations where an experimenter has a limited amount of instances of each random variable. In this case, a straightforward approach would be that of approximating the Bayesian classifier with explicit estimates of f_X and f_Y which, in turn, yields empirically approximated optimum classification boundaries between (possibly disjoint) regions R_X or R_Y. Afterwards, it would be necessary to integrate segments of $P_X f_X$ and $P_Y f_Y$ over the corresponding regions.

Evidently, this practical deployment of a Bayesian classifier is rather indirect (if the goal is just to estimate $\Pr(error)$) and not always viable. On the other hand, effective cardinalities can always be easily estimated trough coincidence counting, and this is indeed our main motivation to propose the very pragmatic method detailed in Section III. Although $Pe^{(2)}$ was shown to be biased, through the former experiments, the proposed method is so simple to implement that it can be a useful additional tool for experimenters with different theoretical backgrounds. Besides, the method is based on the accumulation of any detected coincidence, and can be used as a fast alternative to a rough minimum error estimation, notably in cases with heterogeneous data (*e.g.* categorical and numerical data mixed up). As a matter of fact, coincidence is probably the most fundamental concept in cognition, and the proposed method can be applied wherever coincidence is defined – nothing else is required.

III. A PRAGMATIC METHOD FOR MINIMUM CLASSIFICATION ERROR ESTIMATION THROUGH EFFECTIVE CARDINALITY

To estimate the effective cardinalities, we adapted the quadratic (collision) entropy estimation method originally proposed by S.-K. Ma, in [12], for it is a very simple method that can handle problems where the number of instances is even less than the effective cardinality itself. Ma's method was proposed in the context of Physics and the event *coincidence* was consequently defined in terms of dynamic system states. By contrast, in a broader perspective of pattern recognition, it is helpful to adapt the method by splitting it into three steps, as explained in the sequel.

For a dataset of N_X and N_Y independent instances of X and Y, respectively,

1) Define *coincidence* between instances of X and Y.
2) Estimate effective cardinalities as explained bellow:
    a) Compare all the $T_X = N_X(N_X - 1)/2$ possible pairs of instances of X to find the number of detected coincidences, D_X.
    b) Compare all the $T_Y = N_Y(N_Y - 1)/2$ possible pairs of instances of Y to find the number of detected coincidences, D_Y.
    c) Compare all the $T_{XY} = N_X N_Y$ possible pairs of instances of X *versus* Y to find the number of detected coincidences, D_{XY}.
    d) Estimate the effective cardinalities of virtual sets as $\hat{C}_X = T_X/D_X$, $\hat{C}_Y = T_Y/D_Y$ and $\hat{C}_{XY} = T_{XY}/D_{XY}$.
    e) Estimate the effective cardinality of the intersection between the two virtual sets according to Equation 5, as $\hat{C}_I = \hat{C}_X\hat{C}_Y/\hat{C}_{XY}$.
3) As in Example 1, estimate $\hat{P}e^{(2)} = P_X\frac{\hat{C}_I}{\hat{C}_X}$, if $\frac{P_X}{\hat{C}_X} \leq \frac{P_Y}{\hat{C}_Y}$, or $\hat{P}e^{(2)} = P_Y\frac{\hat{C}_I}{\hat{C}_Y}$, otherwise.

As an illustration of the method, we consider again the classification problem corresponding to Fig. 3, under equal *a priori* probabilities for the classes, to which the optimum

Bayesian classifier yields $\Pr(error) = 0.092$. In step 1 of the method, we define that a coincidence occurs every time two scalar instances differs by less than $\Delta = 0.1$ (see discussion about definition of coincidence for continuous random variables in Subsection III-A). Then, with just 100 instances of each random variable, $N_X = N_Y = 100$, independently drawn for each trial, the proposed method was repeatedly applied 30,000 times, yielding an average estimate of about $\hat{Pe}^{(2)} = 0.088$, and a standard deviation of about 0.028.

In order to preserve the pragmatic leitmotiv of the proposed method, coincidence would be preferably defined according to the expertise of the experimenter. For instance, in a given collection of data, two forms with categorical and/or numerical data may be regarded as two instances of a multivariate signal, and determining whether these two forms are coincident or not may be neither an objective nor a straightforward matter. It is hard to imagine a general definition of coincidence that would fit all practical requirements, regardless of the corresponding domain of application. Moreover, even if a given signal is consistently measured through a well defined scale, we should not expect two coincident measurements in Biology to be necessarily coincident for a physicist, as well. To further discuss this sensitive point, in Subsection III-A, some pragmatic aspects of coincidence definition are highlighted.

A. *On the definition of coincidence*

Although coincidence can be regarded as a fundamental concept in pattern recognition (and in cognition, as a whole), it is difficult to define it mathematically in a consensual manner. Even the idea that coincidence is a Boolean variable (*true* or *false*) can be challenged. For instance, there is an interesting theoretical link between the quadratic entropy estimator in [12] and the Information Theoretical Learning (ITL) framework, started in [13] and further developed in [14], so that coincidences, in ITL approach, can be regarded as a *soft* version of coincidences in Ma's method. More specifically, both approaches aim at directly estimating quadratic entropy from the data samples, without imposing assumptions about random sources distributions, but the kernel (Parzen) based PDF model used in ITL leads to the following equation (transcribed here for reader convenience):

$$\hat{h}^{(2)} = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(x_j - x_i;\sigma)\right), \qquad (12)$$

where continuous degrees of coincidences between pairs of samples, x_i and x_j, are measured through a Gaussian kernel, $G(x_j - x_i;\sigma)$, whose dispersion is parametrized by σ, and are summed up. The resulting sum is divided by the number or pairwise comparisons, N^2. The similarity of this quadratic differential entropy estimator and the method of estimation through coincidences proposed by S.-K. Ma [12] (to be used in statistical mechanics) is striking. By replacing $G(\cdot)$, in Equation 12, with a simpler hard detector (0 or 1) function – also typically used in Parzen models of PDF – the entropy estimators of Equation 12 and that by S.-K. Ma coincide[2].

It is further interesting to note that the nonparametric estimator of entropy in Equation 12 was proposed as early as in 1998, by Dongxin Xu and Jose C. Principe [13], as a robust way to directly extract information from available data (thus avoiding the postulation of parametric PDF models). This corroborate our perception that the gathering all coincidence in a single counter, detected either by a hard or by a soft detector, yields well adapted estimators for limited amounts of data. Nonetheless, as pointed out in [13], it does come with a computational cost, since it is based on the comparison of all available pairs of observations.

Regarding the definition of coincidence itself, deciding between hard or soft coincidence detectors is just the tip of the iceberg, and we do prefer to assume that it is a matter of expertise, with a lot of subjective aspects, rather than an objective theoretical problem. However, for sake of simplicity, *par défault*, we arbitrarily suggest the use of hard of Boolean (*true* or *false*) coincidence definitions. Besides, in cases where the random variable is continuous and defined over a metric space, we also suggest a rule of thumb based on an interesting connection between effective cardinalities and differential entropy. In such cases, a coincidence can be defined, for two instances x_i and x_j of a random variable X, as:

> **Coincidence:** a probabilistic event which is true, for a pair (x_i, x_j), if x_j is found inside a region with *hyper-volume* Δ around x_i . The term hyper-volume can also stand for length, area or volume, depending on the space dimension where X is defined.

In Figure 4, this definition is used with several values of Δ, and 10,000 independent instances were drawn for each of four parametric distributions. The effective cardinality, $C_X^{(2)}$, was re-estimated for each value of Δ, thus allowing the resulting plot of $\log_2(1/\Delta)$ *versus* $\log_2\left(C_X^{(2)}\right)$, where an asymptotic alignment of $\log_2\left(C_X^{(2)}\right)$ with the quadrant diagonal can be noticed. Indeed, as Δ is reduced, for an unlimited amount of instances, it is expected that the vertical distance from the quadrant diagonal to $\log_2\left(C_X^{(2)}\right)$ will tend to the quadratic differential entropy of the corresponding continuous random variable. In other words, the quantity $\hat{h}_X^{(2)} = \log_2(C_X^{(2)}) - \log_2(1/\Delta)$ tends to the quadratic differential entropy, $h_X^{(2)} = -\log_2\left(\int_{-\infty}^{\infty} f_X^2(x)dx\right)$, as Δ tends to zero.

For instance, in the experiments shown in Fig. 4, best estimates and theoretical values of the quadratic differential entropies are compared in Table III-A. This asymptotic behaviour can also be noticed for multivariate variables (*i.e.* dimension $L \geq 1$). Therefore, to keep coincidence definition as simple as possible, the coordinates of x_i may define the center of a hypercube of edge $\xi = \Delta^{1/L}$, so that another instance x_j found inside this hypercube accounts for one coincidence. In other words, for this definition of coincidence, the asymptotic

[2] Ma's method makes $N(N-1)/2$ non-redundant pairwise comparisons, whereas the ITL method makes N^2 redundant ones. Therefore, if $G(\cdot)$ is replaced with the same *hard* coincidence detector used in Ma's approach, both methods provide the same average number of coincidences per comparison.

TABLE I
QUADRATIC DIFFERENTIAL ENTROPIES FOR THE CURVES IN FIGURE 4.

PDF	Theoretical	Estimated
Uniform ($range = 1$)	$\log_2(range) = 0$	0.02
Gaussian ($\sigma = 1$)	$\log_2(2\sigma\sqrt{\pi}) \approx 1.83$	1.84
Exponential ($\lambda = 1$)	$1 - \log_2(\lambda) = 1$	1.03
Laplace ($\lambda = 1$)	$2(1 - \log_2(\lambda)) = 2$	1.99

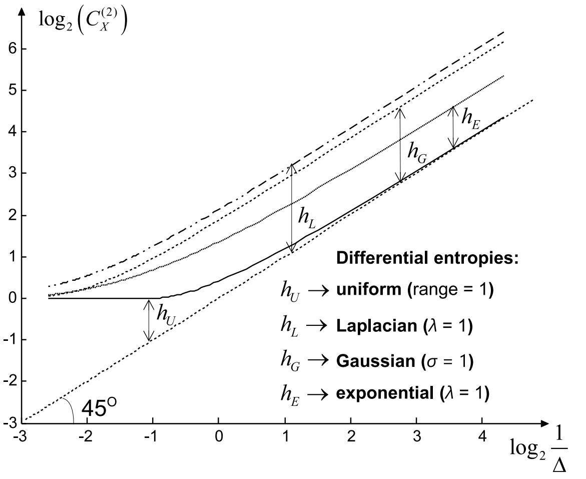


Fig. 4. Differential entropy and the asymptotic behaviour of coincidence counters for continuous variables.

convergence of $h_X^{(2)}$ indicates that the probability density function tends to be uniform inside every hypercube of edge ξ [15]. Thus, a rule of thumb for defining coincidence region for continuous variables can be set according to the following practical iteration:

- Arbitrarily chose a hyper-volume Δ.
- Count all κ coincidences over the whole dataset.
- Double the hyper-volume.
- Recount all coincidences over the whole dataset.
- Accept the hyper-volume Δ as suitable for coincidence definition if the recounted number of coincidences roughly equals 2κ. Otherwise, reduce Δ and move back to step 2.

Unfortunately, too small datasets may not allow this asymptotic behaviour to be observed, mainly if $L \gg 1$, and the definition of coincidence should again depend on the analyst expertise, in which case error probabilities are expected to be as meaningful (of meaningless) as the chosen definition of coincidence.

To illustrate the utility of the proposed method in a case where the optimum classification boundary in nonlinear and two-dimensional, we first gathered independent instances of each random variable, representing two equally probable classes (*i.e.* same *a priori* probabilities), with distributions given by:

$$X_1 \sim \mathcal{N}\left([0,0], \begin{bmatrix} \sigma & 0 \\ 0 & 3 \end{bmatrix}\right), \quad X_2 \sim \mathcal{N}\left([2,0], \begin{bmatrix} \sigma & 0 \\ 0 & 3 \end{bmatrix}\right).$$

to which the Bayesian boundary is clearly a straight vertical line at coordinate 1 of the Cartesian plane, and the minimum classification error can be easily obtained through the integral $\Pr(error) = \frac{1}{\sqrt{2\pi\sigma}} \int_1^\infty \exp\left(\frac{-x^2}{2\sigma^2}\right) dx$.

As an artifice to obtain a perfect nonlinear two-dimensional optimum Bayesian boundary, all instances of X_1 and X_2 are distorted by a nonlinear bijective function, $W = f(X)$, where

$$X = \begin{bmatrix} a \\ b \end{bmatrix} \rightarrow W = \begin{bmatrix} \tanh(0.1(a+b) \\ -\tanh(0.2(a-b)) \end{bmatrix}$$

The resulting distorted clusters of of these points can be seen in Fig. 5, along with the also distorted optimum decision boundary. Since $f(\cdot)$ is a bijective function, the minimum error rate remains unchanged, although the boundary is no longer a straight line, and any classifier whose goal is to minimize the classification error rate would be adjusted, through the available instances of W, to find this optimum boundary.

By contrast, through the proposed method, we defined coincidence with the rule of thumb proposed in this section, with $\xi = 0.01$, and estimates of the minimum error (of a hypothetical well adjusted classifier) were easily obtained. In Fig. 6, these estimates are presented for values of σ ranging from 0.3 to 3.5, with $N = 5000$ independent instances of each class for each value of σ.

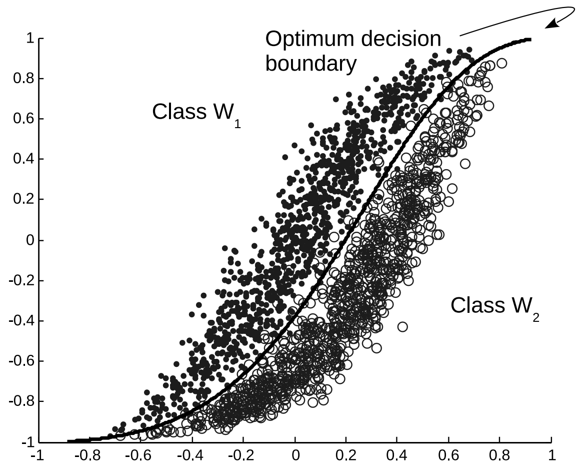


Fig. 5. Nonlinear classification problem example with its optimum decision boundary.

IV. CONCLUSION

A pragmatic method was presented to yield a rough but potentially useful estimate of the minimum probability of classification error, for a two-classes problem. The basic principles that support the method are very simple, namely: cardinality of sets and basic probability concepts, thus providing unbiased error estimates for uniformly distributed random variables.

As for nonuniform random variables, we borrowed the meaningful concept of effective cardinality from [8], and we experimentally showed that the method still provides useful but biased estimates of the minimum error, if the quadratic (collision) entropy is used to obtain effective cardinalities of idealized sets, associated to equally idealized uniform distributions. Estimation biases were found to be either positive or negative, and some interesting cases were studied through numerical simulations.

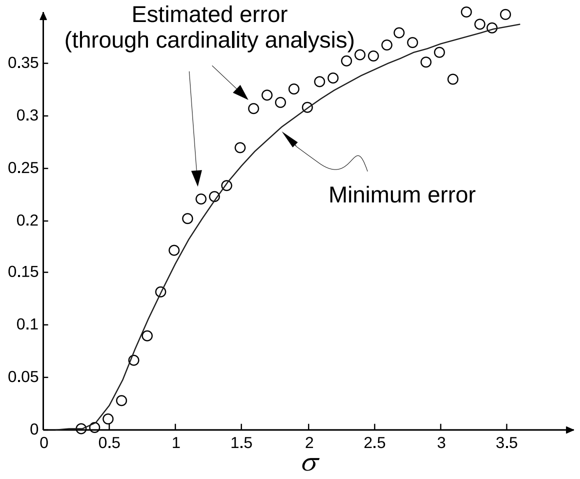


Fig. 6. Decision error for the nonlinear classification problem example – circles correspond to minimum error estimates through effective cardinality comparison, thus without explicit decision boundary estimation.

The proposed method also takes advantage of the fact that quadratic entropy is probably the easiest entropy to be estimated from a limited amount of instances, through simple coincidence counting. For sake of simplicity of use, we divided the method into three steps, where the first one is the definition of the probabilistic event *coincidence*. It was further discussed the importance of the expert to provide meaningful definitions of coincidence, according to the source of the data under analysis. Nonetheless, a rule of thumb for continuous random variable was also provided in Section III-A, based on an expected asymptotic behaviour of differential entropy estimators.

Concerning coincidence definition, yet an interesting theoretical link between the quadratic entropy estimator in [12] and the Information Theoretical Learning (ITL) framework was pointed out here, illustrating that coincidence can even be defined in a continuous scale of values (*i.e. soft* coincidence). This illustration supports the idea that coincidence is a rather difficult to define concept, in spite of its apparent simplicity.

We believe that the effective cardinality concept simplifies the understanding of the practical tools presented in this paper, potentially making it easier to adapt them to experiments done in different domains, regardless of the experimenter fluency in information theory. The pragmatic aspect of the proposed approach may explain, for instance, the connection between the cross-cardinality, $C_{XY}^{(2)}$, estimated through coincidence counting, and the mutual index of coincidence (MIC), used in cryptography [16]. This connection also indicates another potential domain of application for the proposed method, with extension to DNA analysis, where MIC is already in use [17].

Whatever the chosen application, a common aspect of many practical problems related to actual data is the difficulty of applying tools based on entropy estimation to heterogeneous and high-dimensional data. To this concern, the proposed method establishes a net separation between coincidence detection and quantities estimation. This approach doesn't come without risks, for error probabilities are expected to be as meaningful (of meaningless) as the provided definition of coincidence. On the other hand, we also hope that this strategy may simplify the practical use of our method because, in essence, once the event *coincidence* or *collision* is defined by the experimenter, all the other steps are as simple as counting coincidences and making some basic calculations. This strategy also paves the way for application were experimental data is heterogeneous (e.g. mixed categorical and numerical data), since it just demands a proper definition of coincidence for the corresponding heterogeneous data, thus preserving all the other method steps unchanged.

Concerning the practical deployment of the proposed method, the pre-analysis of available data (observed instances), in terms of quantity and quality, is recommended. The main steps of the method rely upon quadratic entropy estimates, therefore reliable such estimates are paramount requirements. For categorical data, as suggested by S.-K. Ma himself [11], if the number of symbols is very small and each symbol is used many times, one would prefer plug in methods (through PMF estimation). Indeed, just as Ma's method for entropy estimation, our method finds its main application potential in cases where the number of instances is much greater than the square-root of the number of symbols, but not enough to yield reliable PMF estimates (e.g. sparse histograms). The theoretical reasons for the robustness of quadratic entropy estimators under data shortage is further studied in the more recent work by Paninski [18].

As for continuous observations in $L-$dimensional observation spaces, even if they are part of a heterogeneous collection of observations (i.e. continuous $L-$dimensional multivariate data coupled to categorical data) they should be separately preprocessed for proper definition of multivariate *coincidence* event. As proposed in [19], in the Section entitled *Setting and testing a coincidence neighbourhood*, if the standard deviations in each dimension of the continuous data are normalized, one may take advantage of an approximative relationship between a first guess for the effective cardinality and the *hyper-volume* Δ. For instance, given a first rough guess for effective cardinality of about $C^{(2)} = 1000$, and $L = 3$, then one may set $\Delta \approx \frac{1.6(3.5)^L}{C^{(2)}} \approx 0.07$, as a first choice for Δ, and then refine it as in Subsection III-A. Note that if the test proposed in Subsection III-A does not hold for any value of Δ, one should conclude that there is not enough data (continuous data) for the entropy estimator to properly work. After categorical and numerical observations are separately tested, they should be jointly tested as well. Practically speaking, the test proposed in Subsection III-A should be applied again, under the definition of coincidence chosen by the experimenter, for the heterogeneous set of observations.

A remaining important issue is to be sure that instances are taken independently, for the coincidence method for entropy estimation assumes it is true. As a practical advice, one should test if it is at least approximately true, mainly in the cases where samples are taken through time from a dynamic phenomenon. In [11], S.-K. Ma drags the reader's attention to what he calls *the relaxation time*, which can be regarded as a phenomenon momentum. Indeed, any dynamic phenomenon may have at least one source of momentum/inertia, regardless whether it is a physical or a social phenomenon (just to mention two) and if samples are sequentially taken without taking it into account, independence between samples may be significantly violated, and entropy estimates, in general, tend to be lowered.

As a final comment on the principles behind this work, as in [8], we also believe that thinking in terms of sets and effective cardinalities – instead of entropy – may induce new interesting ways of thinking about the powerful theoretical background behind the information theory.

Acknowledgment

This work was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to J.M. (grant numbers: 303707/2012-7) and from the Capes PNPD program to J.C.

References

[1] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th Edition, Academic Press, 2008.

[2] C. E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal 27 (3) (1948) 379–423, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x.

[3] R. V. Hartley, Transmission of information, Bell System technical journal 7 (3) (1928) 535–563. doi: 10.1002/j.1538-7305.1928.tb01236.x.

[4] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics (1951) 79–86. doi: 10.1214/aoms/1177729694.

[5] A. Rényi, On measures of entropy and information, in: Fourth Berkeley symposium on mathematical statistics and probability, Vol. 1, 1961, pp. 547–561.

[6] T. van Erven, P. Harremoes, Rényi divergence and kullback-leibler divergence, Information Theory, IEEE Transactions on 60 (7) (2014) 3797–3820. doi: 10.1109/TIT.2014.2320500.

[7] A. Ullah, Entropy, divergence and distance measures with econometric applications, Journal of Statistical Planning and Inference 49 (1) (1996) 137–162. doi: 10.1016/0378-3758(95)00034-8.

[8] M. O. Hill, Diversity and evenness: a unifying notation and its consequences, Ecology 54 (2) (1973) 427–432. doi: 10.2307/1934352.

[9] A. Bialas, W. Czyz, Event by event analysis and entropy of multiparticle systems, Physical Review D 61 (7) (2000) 074021. doi: 10.1103/PhysRevD.61.074021.

[10] I. Nemenman, Coincidences and estimation of entropies of random variables with large cardinalities, Entropy 13 (12) (2011) 2013–2023. doi: 10.3390/e13122013.

[11] S.-K. Ma, Calculation of entropy from data of motion, Journal of Statistical Physics 26 (2) (1981) 221–240. doi: 10.1007/BF01013169.

[12] S. K. Ma, Statistical mechanics, World Scientific Publishing Co Pte Ltd, 1985.

[13] D. Xu, J. C. Principe, Learning from examples with quadratic mutual information, in: Proceedings of 1998 Workshop on Neural Networks for Signal Processing VIII, 1998, pp. 155–164.

[14] J. C. Principe, D. Xu, J. Fisher, Information theoretic learning, in Unsupervised adaptive filtering, Haykin (Ed.), Wiley, (2000) 265–319.

[15] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, New York, NY, USA, 1991.

[16] W. Friedman, The index of coincidence and its applications in cryptography, riverbank publication no. 22, Riverbank Labs., 1922. reprinted in 1987 by Aegean Park press (1922).

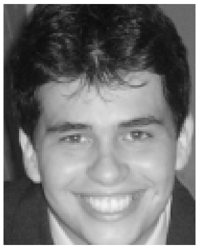
[17] P. Gagniuc, C. Ionescu-Tirgoviste, Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters, BMC genomics 13 (1) (2012) 512. doi: 10.1186/1471-2164-13-512.

[18] L. Paninski, Estimating entropy on m bins given fewer than m samples, IEEE Transactions on Information Theory 50 (9) (2004) 2200–2203. doi: 10.1109/TIT.2004.833360.

[19] J. Montalvão, R. Attux, D. Silva, A pragmatic entropy and differential entropy estimator for small datasets, Journal of Communication and Information Systems 29 (1) (2014) 1–8. doi: 10.14209/jcis.2014.8.

Jugurta Montalvão was born in Aracaju, Brazil, in 1968. He received the title of Electrical Engineer (1992) from the University of Campina Grande (UFPB II), Master in Electrical Engineering (1995) from the University of Campinas (UNICAMP) and Doctor in “Automatique et traitement du signal” (2000) from the University Paris-Sud XI. He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2005. His main research interests are: pattern recognition and signal processing.

Jânio Canuto was born in Maceió, Brazil, in 1984. He received the title of Electrical Engineer (2007) from the Federal University of Sergipe (UFS), M.Sc. in Electrical Engineering (2010) from the State University of Campinas (UNICAMP) and Ph.D. in Computer Science (2014) from Télécom SudParis. He is currently a postdoc fellow at the Department of Computer Science of the Federal University of Sergipe (UFS). His main research interests are: pattern recognition and signal processing.



Elyson Carvalho was born in Arapiraca, Alagoas, Brazil, in 1985. He received the title of Electrical Engineer (2006) from the Federal University of Sergipe (UFS), M.Sc. in Electrical Engineering (2007) from the Federal University of Campina Grande (UFCG) and Ph.D. in Electrical Engineering (2012) from the Federal University of Campina Grande (UFCG). He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2010. His main research interests are: robotics, nonlinear systems, instrumentation and signal processing.