Feature selection with Rényi min-entropy

Marco Romanelli, Catuscia Palamidessi

Feature selection with Rényi min-entropy

Catuscia Palamidessi[1] and Marco Romanelli[1,2]

[1] INRIA, École Polytechnique and University of Paris Saclay, France
[2] Università di Siena, Italy

Abstract. We consider the problem of feature selection, and we propose a new information-theoretic algorithm for ordering the features according to their relevance for classification. The novelty of our proposal consists in adopting Rényi min-entropy instead of the commonly used Shannon entropy. In particular, we adopt a notion of conditional min-entropy that has been recently proposed in the field of security and privacy, and that avoids the anomalies of previously-attempted definitions. This notion is strictly related to the Bayes error, which is a promising property for achieving accuracy in the classification. We evaluate our method on 2 classifiers and 3 datasets, and we show that it compares favorably with the corresponding one based on Shannon entropy.

1 Introduction

The identification of the "best" features for classification is a problem of increasing importance in machine learning. The size of available datasets is becoming larger and larger, both in terms of samples and in terms of features of the samples, and keeping the dimensionality of the data under control is necessary for avoiding an explosion of the training complexity and for the accuracy of the classification [11, 10, 13].

The known methods for reducing the dimensionality can be divided in two categories: those which transform the feature space by reshaping the original features into new ones (*feature extraction*), and those which select a subset of the features (*feature selection*). The second category can in turn be divided in three groups: the *wrapper*, the *embedded*, and the *filter* methods. The last group has the advantage of being classifier-independent, more robust with respect to the risk of overfitting, and more amenable to a principled approach. In particular, several proposals for feature selection have successfully applied concepts and techniques from information theory [2, 20, 9, 15, 4, 19, 3]. The idea is that the smaller is the conditional (aka residual) entropy of the classes given a certain set of features, the more likely the classification of a sample is to be correct. Finding a good set of features corresponds therefore to identifying a set of features, as small as possible, for which such conditional entropy is below a certain threshold.

In this paper, we focus on the filter approach and we propose a new information-theoretical method for feature selection. The novelty of our proposal consists in the use of Rényi min-entropy H_∞ rather than Shannon entropy. As far as we know, all the previous proposals are based on Shannon entropy, with the notable exception of [8] who considered the Rényi entropies. However [8] reported experimental results only on other orders of Rényi entropies. We will discuss more that work in Section 5.

One reason why Rényi min-entropy has not been used more widely may be that what is needed for feature selection is its conditional version, and Rényi did not define it. There have been various attempts to define the conditional min-entropy, but they were unsuccessful because they led to anomalies. For instance, [5] defined the conditional min-entropy of X given Y along the lines of conditional Shannon entropy, namely as the expected value of the entropy of X for each given value of Y. Such definition, however, violates the *data processing inequality*. In particular knowing the value of Y could increase the entropy of X rather than diminishing it.

Recently, however, some advances in the fields of security and privacy have revived the interest for the Rényi min-entropy. The reason is that it models a basic notion of attacker: the (*one-try*) *eavesdropper*. Such attacker tries to infer a secret (e.g., a key, a password, etc.) from the observable behavior of the system, with the limitation that he can try only once. Naturally, a rational attacker will try to minimize the probability of error, so he will pick the secret with the highest probability given what he has observed. Note the similarity with the classification problem, where we choose a class on the basis of the observed feautures, trying to minimize the probability of mis-classification.

Driven by the motivation of providing an information-theoretic interpretation of the eavesdropper operational behavior, [17] proposed a definition of conditional min-entropy $H_\infty(X|Y)$ which is consistent with the rest of the theory, models all the expected properties of an eavesdropper, and corresponds closely to the Bayes risk of guessing the wrong secret. (The formal definition of $H_\infty(X|Y)$ will be given in Section 2.) It is then natural to investigate whether this new notion can be successfully applied also to the problem of feature selection.

We could state the problem of feature selection as finding a minimum-size subset S of the whole set of features F such that the min-entropy $H_\infty(C|S)$ of the classification C given S is below a given threshold. Because of the correspondence with the Bayes risk, this would mean that the set S is optimal (i.e, minimal) among the subsets for which the Bayes classifier achieves the desired level of accuracy. However, is that the construction of such an optimal S would be NP-hard. This is not due to the kind of entropy that we choose, but simply to the fact that it is a combinatorial problem. In [12] it was shown that the problem of feature selection can be modeled as search problem on a decision tree, and it was argued that finding the optimal subtree which is able to cover F is an NP-hard problem. The same intractability was claimed in [10] with respect to wrappers and embedded methods, on the basis of the proof of [1].

We then adopt a greedy strategy to approximate the minimal subset of features: following [4] and [19], we construct a sequence of subsets $S^0, S^1, \ldots, S^t, \ldots$, where $S^0 = \emptyset$ and at each subsequent step S^{t+1} is obtained from S^t by adding the next feature in order of relevance for the classification, taking into account the ones already selected. In other words, we select the feature f such that $H_\infty(C|S^t \cup \{f\})$ is minimal, and we define S^{t+1} as $S^t \cup \{f\}$. The construction of this series should be interleaved with a test on the accuracy of the intended classifier(s): when we obtain an S^T that achieves the desired level of accuracy, we can stop. The difference with respect to [4] and [19] is that the relevance is measured by Rényi min-entropy rather than Shannon entropy.

Note that, because of the relation between the conditional min-entropy and the Bayes risk, our method is *locally optimal*. Namely, for any other possible feature f'

(including the one that would be selected using Shannon entropy), the set S^{t+1} is at least as good as $S^t \cup \{f'\}$ in terms of accuracy of the Bayes classifier (the ideal classifier giving the best accuracy). This does not necessarily mean that the set S^T is the smallest one: since we are not making an exhaustive search on all possible subsets of F, and we add the features one by one, we may not find the "shortest path" to achieve sufficient accuracy. The same applies to the analogous algorithms based on Shannon entropy. Hence we have no guarantee that our method is better than that of [4] and [19], nor vice versa. In the experiments we have performed, however, our method outperforms almost always the one based on Shannon entropy (cfr. Section 4).

2 Preliminaries

In this section we briefly review some basic notions from probability and information theory. We refer to [7] for more details.

Let X, Y be discrete random variables with respectively n and m possible values: $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$. Let $p_X(\cdot)$ and $p_Y(\cdot)$ indicate the probability distribution associated to X and Y respectively, and let $p_{Y,X}(\cdot,\cdot)$ and $p_{Y|X}(\cdot|\cdot)$ indicate the joint and the conditional probability distributions, respectively. Namely, $p_{Y,X}(x,y)$ represents the probability that $X = x$ and $Y = y$, while $p_{Y|X}(y|x)$ represents the probability that $Y = y$ given that $X = x$. For simplicity, when clear from the context, we will omit the subscript, and write for instance $p(x)$ instead of $p_X(x)$.

Conditional and joint probabilities are related by the chain rule $p(x,y) = p(x)\,p(y|x)$, from which (by the commutativity of $p(x,y)$) we can derive the Bayes theorem:

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} \tag{1}$$

2.1 Rényi entropies, Shannon entropy, and mutual information

The Rényi entropies ([16]) are a family of functions representing the uncertainty associated to a random variable. Each Rényi entropy is characterized by a non-negative real number α (order), with $\alpha \neq 1$, and is defined as

$$H_\alpha(X) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log \left(\sum_i p(x_i)^\alpha \right)$$

If $p(\cdot)$ is uniform then all the Rényi entropies are equal to $\log |X|$. Otherwise they are weakly decreasing in α. Shannon and min-entropy are particular cases:

$\alpha \to 1$	$H_1(X) = -\sum_x p(x) \log p(x)$	Shannon entropy
$\alpha \to \infty$	$H_\infty(X) = -\log \max_x p(x)$	min-entropy

Shannon *conditional entropy* of X given Y represents the average residual entropy of X once the value of Y is known, and it is defined as

$$H_1(Y|X) \stackrel{\text{def}}{=} \sum_{xy} p(x,y) \log p(x|y) = H_1(X,Y) - H_1(Y) \tag{2}$$

where $H_1(X, Y)$ represents the entropy of the intersection of X and Y.

Shannon *mutual information* of X and Y represents the correlation of information between X and Y, and it is defined as

$$I_1(X;Y) \stackrel{\text{def}}{=} H_1(X) - H_1(X|Y) = H_1(X) + H_1(Y) - H_1(X,Y) \tag{3}$$

It is possible to show that $I_1(X;Y) \geq 0$, with $I_1(X;Y) = 0$ iff X and Y are independent, and that $I_1(X;Y) = I_1(Y;X)$.

Rényi did not define conditional entropy and mutual information for generic α. Only recently, thanks to its applications in security and privacy, there has been interest in "completing the theory", and [17] proposed the following definition for the conditional min-entropy:

$$H_\infty(X|Y) \stackrel{\text{def}}{=} -\log \sum_y \max_x (p(y|x)p(x)) \tag{4}$$

It is possible to show that this definition closely corresponds to the Bayes risk, i.e., the expected error when we try to guess the exact value of X, once we know that of Y, formally the Bayes risk of X given Y is defined as:

$$\mathcal{B}(X \mid Y) \stackrel{\text{def}}{=} 1 - \sum_y p(y) \max_x p(x|y) \tag{5}$$

The "mutual information" is defined as:

$$I_\infty(X;Y) \stackrel{\text{def}}{=} H_\infty(X) - H_\infty(X|Y) \tag{6}$$

It is possible to show that $I_\infty(X;Y) \geq 0$, and that $I_\infty(X;Y) = 0$ if X and Y are independent (the reverse is not necessarily true). It is important to note that, contrary to Shannon mutual information, I_∞ is not symmetric.

The conditional mutual information is defined as

$$I_\infty(X;Y|Z) \stackrel{\text{def}}{=} H_\infty(X|Z) - H_\infty(X|Y,Z) \tag{7}$$

and analogously for Shannon conditional mutual information.

3 Our proposed algorithm

Let F be the set of features at our disposal, and let C be the set of classes. Our algorithm is based on forward feature selection and dependency maximization: it constructs a monotonically increasing sequence $\{S^t\}_{t\geq 0}$ of subsets of F, and, at each step, the subset S^{t+1} is obtained from S^t by adding the next feature in order of importance (i.e., the informative contribution to classification), taking into account the information already provided by S^t. The measure of the "order of importance" is based on conditional min-entropy. The construction of the sequence is assumed to be done interactively with a test on the accuracy achieved by the current subset, using one or more classifiers. This test will provide the stopping condition: once we obtain the desired level of accuracy, the algorithm stops and gives as result the current subset S^T. Of course, achieving a level of accuracy $1 - \varepsilon$ is only possible if $\mathcal{B}(C \mid F) \leq \varepsilon$.

Definition 1. *The series* $\{S^t\}_{t\geq 0}$ *and* $\{f^t\}_{t\geq 1}$ *are inductively defined as follows:*

$$\begin{aligned} S^0 &\stackrel{\text{def}}{=} \emptyset \\ f^{t+1} &\stackrel{\text{def}}{=} \operatorname{argmin}_{f\in F\setminus S^t} H_\infty(C \mid f, S^t) \\ S^{t+1} &\stackrel{\text{def}}{=} S^t \cup \{f^{t+1}\} \end{aligned}$$

The algorithms in [4] and [19] are analogous, except that they use Shannon entropy. They also define f^{t+1} based on the maximization of mutual information instead of the minimization of conditional entropy, but this is irrelevant. In fact

$$I_\infty(C; f \mid S^t) = H_\infty(C \mid S^t) - H_\infty(C \mid f, S^t)$$

hence maximizing $I_\infty(C; f \mid S^t)$ with respect to f is the same as minimizing $H_\infty(C \mid f, S^t)$ with respect to f. The same holds for Shannon entropy.

Our algorithm is locally optimal, in the sense stated by the following proposition:

Proposition 1. *At every step, the set* S^{t+1} *minimizes the Bayes risk of the classification among those which are of the form* $S^t \cup \{f\}$*, namely:*

$$\forall f \in F \quad \mathcal{B}(C \mid S^{t+1}) \leq \mathcal{B}(C \mid S^t \cup \{f\})$$

Proof. Let $\boldsymbol{v}, v, v'$ represent generic value tuples and values of S^t, f and f^{t+1}, respectively. Let c represent the generic value of C. By definition, $H_\infty(C \mid S^{t+1}) \leq H_\infty(C \mid S^t \cup \{f\})$, for every $f \in F$. From (4) we then obtain

$$\sum_{\boldsymbol{v},v} \max_c (p(\boldsymbol{v}, v|c)p(c)) \leq \sum_{\boldsymbol{v},v'} \max_c (p(\boldsymbol{v}, v'|c)p(c))$$

Using the Bayes theorem (1), we get

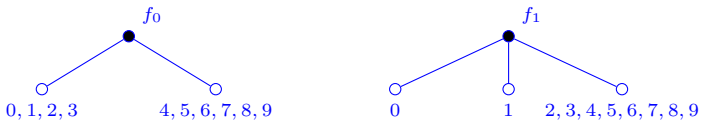
$$\sum_{\boldsymbol{v},v} p(\boldsymbol{v}, v) \max_c p(c|\boldsymbol{v}) \leq \sum_{\boldsymbol{v},v'} p(\boldsymbol{v}, v') \max_c p(\boldsymbol{v}, v'|c)$$

Then, from the definition (5) we deduce

$$\mathcal{B}(C \mid S^t \cup \{f^{t+1}\}) \leq \mathcal{B}(C \mid S^t \cup \{f\})$$

In the following sections we analyze some extended examples to illustrate how the algorithm works, and also compare the resulting sequence with the one produced by the algorithm of [4] and [19].

3.1 An example in which Rényi min-entropy gives a better feature selection than Shannon entropy



(a) Selection with Shannon. (b) Selection with Rényi.

Fig. 2. Classes separation after the selection of the first feature.

Let us consider the dataset in Fig. 1, containing ten records labeled each by a different class, and characterized by six features (columns $f_1, \ldots, f_5$). We note that column f_0 separates the classes in two sets of four and six elements respectively, while all the other columns are characterized by having two values, each of which univocally identify one class, while the third value is associated to all the remaining classes. For instance, in column f_1 value A univocally identifies the record of class 0, value B univocally identifies the record of class 1, and all the other records have the same value along that column, i.e. C.

Class	f_0	f_1	f_2	f_3	f_4	f_5
0	A	C	F	I	L	O
1	A	D	F	I	L	O
2	A	E	G	I	L	O
3	A	E	H	I	L	O
4	B	E	F	J	L	O
5	B	E	F	K	L	O
6	B	E	F	I	M	O
7	B	E	F	I	N	O
8	B	E	F	I	L	P
9	B	E	F	I	L	Q

Fig. 1. The dataset

The last five features combined completely identify all classes, without the need of the first one. On the other hand, the last five features are also necessary to completely identify all classes, they cannot be replaced by f_0 (i.e., f_0 with whatever combination of other four or less features is not sufficient to completely identify all classes.) In fact, each pair of records which are separated by one of the features $f_1, \ldots, f_5$, have the same value in column f_0.

If we apply the discussed feature selection method and we look for the feature that minimizes $H(Class|f_i)$ for $i \in \{0, \ldots, 5\}$ we obtain that:

- The first feature selected by the Shannon entropy is f_0, in fact $H_1(Class|f_0) \approx 2.35$ and $H_1(Class|f_{\neq 0}) = 2.4$. (The notation $f_{\neq 0}$ stands for any of the f_i's except f_0.) In general, indeed, with Shannon entropy the method tends to choose a feature which splits the classes in a way as balanced as possible. The situation after the selection of the feature f_0 is shown in Fig. 2(a).
- The first feature selected by the Rényi min entropy is either f_1 or f_2 or f_3 or f_4 or f_5, in fact $H_\infty(Class|f_0) \approx 2.32$ and $H_\infty(Class|f_{\neq 0}) \approx 1.74$. In general, indeed, with Rényi min-entropy the method tends to choose a feature which divides the classes in as many sets as possible. The situation after the selection of the feature f_1 is shown in Fig. 2(b).

Going ahead with the algorithm, with Shannon entropy we will select one by one all the other features, and as already discussed we will need all of them to completely identify all classes. Hence at the end the method with Shannon entropy will return all the six features (to achieve perfect classification). On the other hand, with Rényi min entropy we will select all the remaining features except f_0 to obtain the perfect discrimination. In fact, at any stage the selection of f_0 would allow to split the remaining classes in at most two sets, while any other feature not yet considered will split the remaining

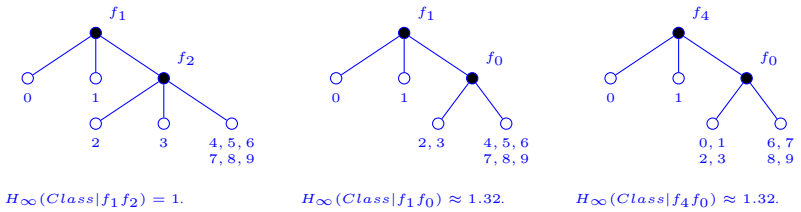


Fig. 3. Selection of the second feature with Rényi.

classes in three sets. As already hinted, with Rényi we choose the feature that allows to split the remaining classes in the highest number of sets, hence we never select f_0. For instance, if we have already selected f_1, we have $H_\infty(Class|f_1 f_0) \approx 1.32$ while $H_\infty(Class|f_1 f_{\neq 0}) = 1$. If we have already selected f_4, we have $H_\infty(Class|f_4 f_0) \approx 1.32$ while $H_\infty(Class|f_4 f_{\neq 0}) = 1$. See Fig. 3.

At the end, the selection of features using Rényi entropy will determine the progressive splitting represented in Fig. 4. The order of selection is not important: this particular example is conceived so that the features $f_1, \ldots, f_5$ can be selected in any order, the residual entropy is always the same.

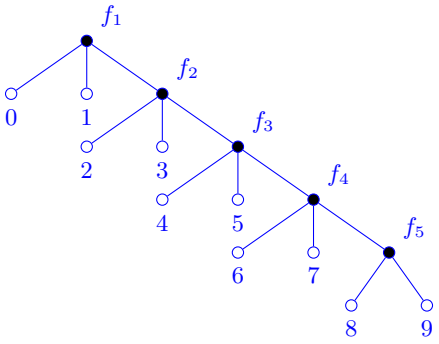


Fig. 4. Sequence of class splitting with Rényi.

Discussion It is easy to see that, in this example, the algorithm based on Rényi min-entropy gives a better result not only at the end, but also at each step of the process. Namely, at step t (cfr. Definition 1) the set S^t of features selected with Rényi min-entropy gives a better classification (i.e., more accurate) than the set S'^t that would be selected using Shannon entropy. More precisely, we have $\mathcal{B}(C \mid S^t) < \mathcal{B}(C \mid S'^t)$. In fact, as discussed above the set S'^t contains necessarily the feature f_0, while S^t does not. Let S^{t-1} be the set of features selected at previous step with Rényi min-entropy, and f^t the feature selected at step t (namely, $S^{t-1} = S^t \setminus \{f^t\}$). As argued above, the order of selection of the features $f_1, \ldots, f_5$ is irrelevant, hence we have $\mathcal{B}(C \mid S^{t-1}) = \mathcal{B}(C \mid S'^t \setminus \{f_0\})$ and the algorithm *could* equivalently have selected $S'^t \setminus \{f_0\}$. As argued above, the next feature to be selected, with Rényi, must be different from f_0. Hence by Proposition 1, and by the fact that the order of selection of $f_1, \ldots, f_5$ is irrelevant, we have: $\mathcal{B}(C \mid S^t) = \mathcal{B}(C \mid (S'^t \setminus \{f_0\}) \cup \{f^t\}) < \mathcal{B}(C \mid S'^t)$.

As a general observation, we can see that the method with Shannon tends to select the feature that divides the classes in sets (one for each value of the feature) as balanced as possible, while our method tends to select the feature that divides the classes in as many sets as possible, regardless of the sets being balanced or not. In general, both Shannon-based and Rényi-based methods try to minimize the height of the tree representing the process of the splitting of the classes, but the first does it by trying to

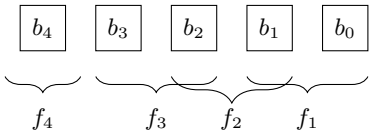
produce a tree *as balanced as possible*, while the second try to do it by producing a tree *as wide as possible*. Which of the method is best, it depends on the correlation of the features. Shannon works better when there are enough uncorrelated (or not much correlated) features, so that the tree can be kept balanced while being constructed. Next section shows an example of such situation. Rényi, on the contrary, is not so sensitive to correlation and can work well also when the features are highly correlated, as it was the case in the example of this section.

The experimental results in Section 4 show that, at least in the cases we have experimented with, our method based on Rényi min-entropy outperforms the one based on Shannon entropy. In general however the two methods are incomparable, and perhaps a good practice would be to construct both sequences at the same time, so to obtain the best result of the two.

3.2 An example in which Shannon entropy eventually gives a better feature selection than Rényi min-entropy

Consider a dataset containing samples equally distributed among 32 classes, indexed from 0 to 31. As usual, we denote by C the random variable ranging over the classes. Assume that the data have 8 features divided in 2 types F and F', each of which consisting of 4 features denoted as follows: $F = \{f_1, f_2, f_3, f_4\}$ and $F' = \{f'_1, f'_2, f'_3, f'_4\}$. Thus the total set of features is $F \cup F'$. The features f_i, f'_i are also regarded as random variables.

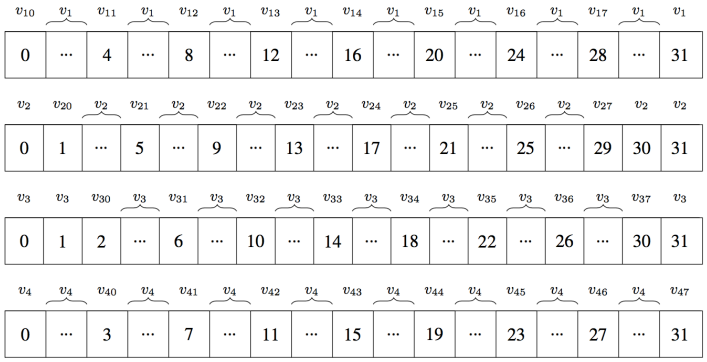
Assume that the features of type F, and their relation with C are as follows: Given the binary representation of the classes $b_4b_3b_2b_1b_0$, f_1 consists of the bits b_1b_0, f_2 consists of the bits b_2b_1, f_3 consists of the bits b_3b_2, and f_4 consists of the bit b_4 only. The situation is represented by the figure below.



Concerning the features of type F', assume that each of them can range in a set of 9 values, and that these sets are mutually disjoint. We use the following notation:

$$\text{for } 1 \leq i \leq 4 \quad f'_i = \{v_i\} \cup \{v_{ij} \mid 0 \leq j \leq 7\}$$

The association between the features of F' and the classes is represented in the following figure:



All the v_{ij} values allow to identify exactly one class. For instance, if in a certain sample the f'_1 value is v_{10}, it means that the sample can be univocally classified as belonging to class 0. If the f'_2 value is v_{22}, it belongs to class 9, etc. As for the v_i's, each of them is associated to a set of 24 classes, with uniform distribution. For instance, the value v_1 is associated to classes $1, 2, 3, 5, \ldots, 29, 30, 31$.

Let us select the first feature f in order of importance for the classification. We will consider at the same time our method, based on Rényi min-entropy, and the method of [4] and [19]. In both cases, the aim is to choose f such that it maximizes the mutual information $I_\alpha = H_\alpha(C) - H_\alpha(C \mid f)$, or equivalently, that minimizes the conditional entropy $H_\alpha(C \mid f)$, where α is the entropy parameter ($\alpha = 1$ for Shannon, $\alpha = \infty$ for Rényi min-entropy).

For the features of type F we have:

$$H_1(C \mid f_1) = H_1(C \mid f_2) = H_1(C \mid f_3) = 3$$

while

$$H_1(C \mid f_4) = 4 > 3$$

On the other hand, with respect to the features of type F', we have

$$H_1(C \mid f'_1) = H_1(C \mid f'_2) = H_1(C \mid f'_3) = H_1(C \mid f'_4) \approx 3.439 > 3$$

So, the first feature selected using Shannon entropy would be f_1, f_2 or f_3. Let us assume that we pick f_1. Hence with Shannon the first set of the series is $S_1^1 = \{f_1\}$.
Let us now consider our method based on Rényi min-entropy. For the F feature the conditional entropy is the same as for Shannon:

$$H_\infty(C \mid f_1) = H_\infty(C \mid f_2) = H_\infty(C \mid f_3) = 3$$

$$H_\infty(C \mid f_4) = 4 > 3$$

But for the F' features Rényi min-entropy gives a different value. In fact we get:

$$H_\infty(C \mid f'_1) = H_\infty(C \mid f'_2) = H_\infty(C \mid f'_3) = H_\infty(C \mid f'_4) \approx 1.83 < 2$$

So, the first feature selected using Rényi min entropy would be f'_1, f'_2, f'_3, or f'_4. Let us assume that we pick f'_1. Hence with our method the first set of the series is $S^1_\infty = \{f'_1\}$.

For the selection of the second feature, we have

$$H_1(C \mid f_1, f_2) = H_1(C \mid f_1, f_4) = 2$$

while

$$H_1(C \mid f_1, f_3) = 1$$

$$H_1(C \mid f_1, f'_1) = H_1(C \mid f_1, f'_2) = H_1(C \mid f_1, f'_3) = H_1(C \mid f_1, f'_4) > 2$$

Hence the second feature with Shannon can only be f_3, and thus the second set in the sequence is $S^2_1 = \{f_1, f_3\}$.

With Rényi min-entropy we have:

$$H_\infty(C \mid f'_1, f_4) > H_\infty(C \mid f'_1, f_i) \text{ for } i = 1, 2, 3$$

$$H_\infty(C \mid f'_1, f_i) > H_\infty(C \mid f'_1, f'_j) \text{ for } i = 1, 2, 3 \text{ and } j = 2, 3, 4$$

Hence with our method the second feature will be f'_2, f'_3, or f'_4. Let us assume that we choose f'_2. Thus the second set of the series is $S^2_\infty = \{f'_1, f'_2\}$.

Continuing our example, we can see that $S^3_1 = \{f_1, f_3, f_4\}$, and $S^3_\infty = \{f'_1, f'_2, f'_i\}$ where i can be, equivalently, 3 or 4. At this point the method with Shannon can stop, since the residual Shannon entropy of the classification is $H_1(C \mid S^3_1) = 0$, and also the Bayes risk is $\mathcal{B}(C \mid S^3_1) = 0$, which is the optimal situation in the sense that the classification is completely accurate. S^3_∞ on the contrary is not enough to give a completely accurate classification, for that we have to make a further step. We can see that $S^4_\infty = F'$, and finally we have $H_\infty(C \mid S^4_\infty) = 0$.

Thus in this particular example we have that for small values of the threshold on the accuracy our method gives better results. On the other hand, if we want to achieve perfect accuracy (threshold 0) Shannon gives better results. This is not always the case however: for example, if we consider the initial set of features to be $F \cup F'$, with $F = \{f_1\}$, with f_1 and F' defined as before, we have that our method achieves perfect accuracy at step 4, giving as result $S^4_\infty = F'$ as before, while with Shannon the method would still select as first feature f_1, and would achieve perfect accuracy only at step 5, giving as result $S^5_1 = F \cup F'$.

4 Evaluation

In this section we evaluate the method for feature selection that we have proposed, and we compare it with the one based on Shannon entropy by [4] and [19].

To evaluate the effect of feature selection, some classification methods have to be trained and tested on the selected data. We decided to use two different methods so to avoid the dependency of the result on a particular algorithm. We chose two widely used classifiers, which in many cases represent the state of the art in the machine learning field, namely Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

Even though the two methods are very different, they have in common that their efficiency is highly dependent on the choice of certain parameters. Therefore, it is worth

spending some effort to identify the best values. Furthermore, we should take into account that the particular paradigm of SVM we chose only needs 2 parameters to be set, while for ANN the number of parameters increases (at least 4).

It is very important to choose values as robust as possible for the parameters. It goes without saying that the strategy used to pick the best parameter setting should be the same for both Shannon entropy and Rényi min-entropy. On the other hand for SVM and ANN we used two different hyper-parameter tuning algorithms, given that the number and the nature of the parameters to be tuned for those classifiers is different.

In the case of SVM we tuned the cost parameter of the objective function for margin maximization (*C-SVM*) and the parameter which models the shape of the RBF kernel's bell curve (γ). Grid-search and Random-search are quite time demanding algorithms for the hyper-parameter tuning task but they're also widely used and referenced in literature when it comes to SVM. Following the guidelines in [6] and [14], we decided to use Grid-search, which is quite suitable when we have to deal with only two parameters. In particular we performed Grid-search including a 10 folds CV step.

Things are different with ANN because many more parameters are involved and some of them change the topology of the network itself. Among the various strategies to attack this problem we picked Bayesian Optimization [18]. This algorithm combines steps of extensive search for a limited number of settings before inferring via Gaussian Processes (GP) which is the best setting to try next (with respect to the mean and variance and compared to the best result obtained in the last iteration of the algorithm). In particular we tried to fit the best model by optimizing the following parameters:

- number of hidden layers
- number of hidden neurons in each layer
- learning rate for the gradient descent algorithm
- size of batches to update the weight on network connections
- number of learning epochs

To this purpose, we included in the pipeline of our code the *Spearmint* Bayesian optimization codebase. Spearmint, whose theoretical bases are explained in [18], calls repeatedly an objective function to be optimized. In our case the objective function contained some *tensorflow* machine learning code which run a 10 folds CV over a dataset and the objective was to maximize the accuracy of validation. The idea was to obtain a model able to generalize as much as possible using only the selected features before testing on a dataset which had never been seen before.

We had to decide the stopping criterion, which is not provided by *Spearmint* itself. For the sake of simplicity we decided to run it for a time lapse which has empirically been proven to be sufficient in order to obtain results meaningful for comparison. A possible improvement would be to keep running the same test (with the same number of features) for a certain amount of time without resetting the computation history of the package and only stop testing a particular configuration if the same results is output as the best for k iterations in a row (for a given k).

Another factor, not directly connected to the different performances obtained with different entropies, but which is of the highest importance for the optimization of ANN, is the choice of the activation functions for the layers of neurons. In our work we have been using ReLU activation function for all layers because it is well known as a function

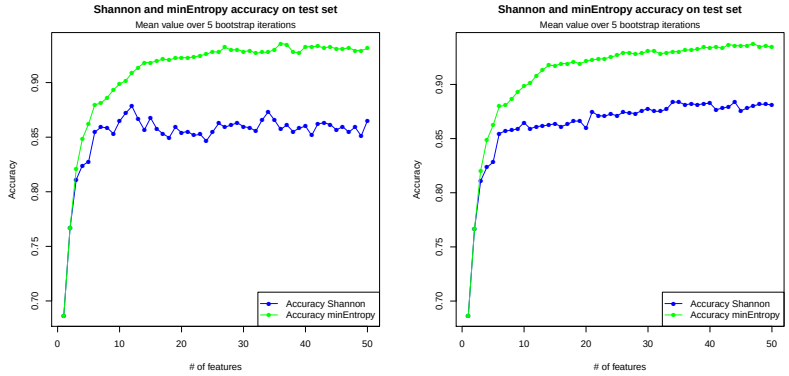


Fig. 5. Accuracy of the ANN and SVM classifiers on the BASEHOCK dataset

which works well for this aim, is quite easy to compute (the only operation involved is the max) and avoids the sigmoid saturation issue.

4.1 Experiments

As already stated, at the i-th step of the feature selection algorithm we consider all the features which have already been selected in the previous $i-1$ step(s). For the sake of limiting the execution time, we decided to consider only the first 50 selected features with both metrics. We tried our pipeline on the following datasets:

- BASEHOCK dataset: 1993 instances, 4862 features, 2 classes. This dataset has been obtained from the 20 newsgroup original dataset.
- SEMEION dataset: 1593 instances, 256 features, 10 classes. This is a dataset with encoding of hand written characters.
- GISETTE dataset: 6000 instances, 5000 features, 2 classes. This is the discretized version of the NIPS 2003 dataset which can be downloaded from the site of Professor Gavin Brown, Manchester University.

We implemented a bootstrap procedure (5 iterations on each dataset) to shuffle data and make sure that the results do not depend on the particular split between training, validation and test set. Each one of the 5 bootstrap iterations is a new and unrelated experimental run. For each one of them a different training-test sets split was taken into account. Features were selected analyzing the training set (the test set has never been taken into account for this part of the work). After the feature selection was executed according to both Shannon and Rényi min-entropy, we considered all the selected features adding one at each time. So, for each bootstrap iteration we had 50 steps, and in each step we added one of the selected features, we performed hyper-parameter tuning with 10 folds CV, we trained the model with the best parameters on the whole training set and we tested it on the test set (which the model had never seen so far). This procedure was performed both for SVM and ANN.

We computed the average performances over the 5 iterations and we got the results showed in Figures 5, 6, and 7. As we can see, in all cases the feature selection method

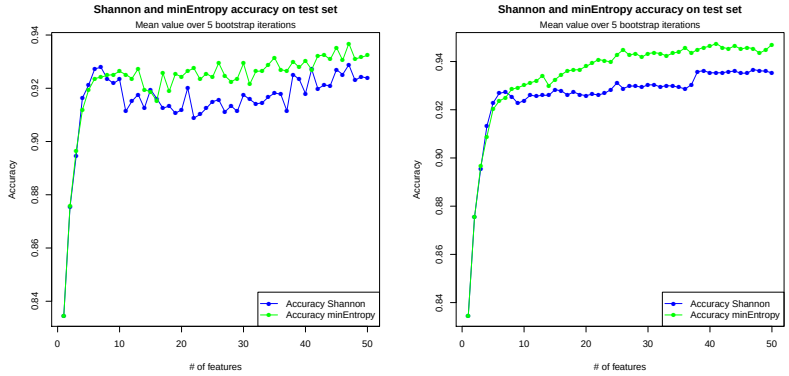


Fig. 6. Accuracy of the ANN and SVM classifiers on the GISETTE dataset

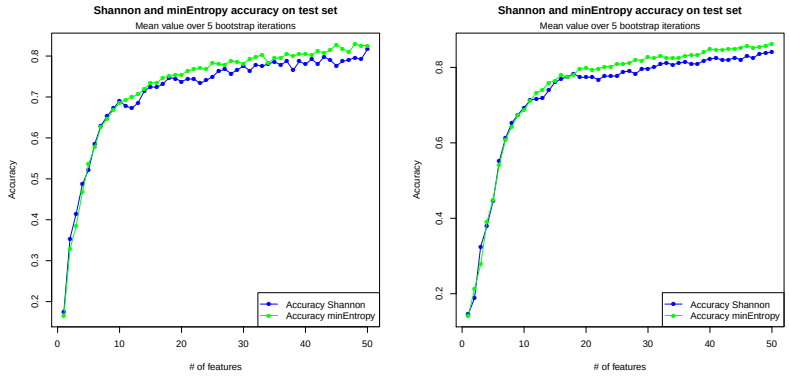


Fig. 7. Accuracy of the ANN and SVM classifiers on the SEMEION dataset

using Rényi min-entropy usually gave better results than Shannon, especially with the BASEHOCK dataset.

5 Related works

In the last two decades, thanks to the growing interest in machine learning, many methods have been setup to tackle the feature reduction problem. In this section we discuss those closely related to our work, namely those which are based on information theory. For a more complete overview we refer to [3], [19] and [4].

The approach most related to our proposal is that of [4] and [19]. We have already discussed and compared their method with ours in the technical body of this paper.

As far as we know, Rényi min-entropy has only been used, in the context of feature selection, by [8]. In the experiments they only show results for other Rényi entropies, not for the min one. But they mention also Rényi min-entropy, though. The definition of conditional min-entropy they consider, however, is that of [5]. This notion, as we have already mentioned, has unnatural consequences. In particular, under this definition, a

feature may increase the entropy of the classification instead of decreasing it. It is clear, therefore, that basing a method on this notion of entropy could lead to strange results.

Two key concepts that have been widely used are *relevance* and *redundancy*. Relevance refers to the importance for the classification of the feature under consideration f^t, and it is in general modeled as $I(C; f^t)$, where I represents Shannon mutual information. Redundancy represents how much the information of f^t is already covered by S. It is often modeled as $I(f^{\mathrm{t}}, S)$. In general, we want to maximize relevance and minimize redundancy.

One of the first algorithms ever implemented was proposed by [2] and it is called MIFS algorithm. This algorithm is based on a greedy strategy. At the first iteration step it selects the feature $f^1 = \mathrm{argmax}_{f_i \in F} I(C; f_i)$, and at step t it selects $f^{\mathrm{t}} = \mathrm{argmax}_{f_i \in F \setminus S^{\mathrm{t\text{-}1}}} [I(C, f_i) - \beta \sum_{f_s \in S^{\mathrm{t\text{-}1}}} I(f_i, f_s)]$ where β is a parameter that controls the weight of the redundancy part.

The mRMR approach (redundancy minimization and relevance maximization) proposed by [15] is based on the same strategy as MIFS. However the redundancy term is now substituted by its mean over the $\mid S \mid$ elements of subset S so to avoid its value to grow when new attributes are selected.

In both cases, if relevance outgrows redundancy, it might happen that many features highly correlated and so highly redundant can still be selected. Moreover, a common issue with these two methods is that they do not take into account the conditional mutual information $I(C, f^{\mathrm{t}} \mid S)$ for the choice of f^{t}, the next feature to be selected.

More recent algorithms involve the ideas of joint mutual entropy $I(C; f_i, S)$ (JMI, [3]) and conditional mutual entropy $I(C; f_i \mid S)$ (CMI, [9]) The deductive step for choosing the next feature for [3] is $f^{\mathrm{t}} = \mathrm{argmax}_{f_i \in F \setminus S^{\mathrm{t\text{-}1}}} \{min_{f_s \in S^{\mathrm{t\text{-}1}}} I(C; f_i, f_s)\}$, while for [9] is $f^{\mathrm{t}} = \mathrm{argmax}_{f_i \in F \setminus S^{\mathrm{t\text{-}1}}} \{min_{f_s \in S^{\mathrm{t\text{-}1}}} I(C; f_i \mid f_s)\}$. In both cases the already selected features are taken into account one by one when compared to the new feature to be selected f^{t}. The correlation between JMI and CMI is easy to prove [20]:

$$I(C; f_i, S) = H(C) - H(C \mid S) + H(C \mid S) - H(C \mid S) = I(C; S) + I(C; f_i \mid S)$$

References

1. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theor. Comput. Sci **209**(1-2), 237–260 (1998)
2. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks **5**(4), 537–550 (1994)
3. Bennasar, M., Hicks, Y., Setchi, R.: Feature selection using joint mutual information maximisation. Expert Syst. Appl **42**(22), 8520–8532 (2015)
4. Brown, G., Pocock, A.C., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research **13**, 27–66 (2012), http://dl.acm.org/citation.cfm?id=2188387
5. Cachin, C.: Entropy Measures and Unconditional Security in Cryptography. Ph.D. thesis, ETH Zurich (1997), reprint as vol. 1 of *ETH Series in Information Security and Cryptography*, ISBN 3-89649-185-7, Hartung-Gorre Verlag, Konstanz, 1997
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. J. Wiley & Sons, Inc. (1991)
8. Endo, T., Kudo, M.: Weighted naïve bayes classifiers by renyi entropy. In: Ruiz-Shulcloper, J., di Baja, G.S. (eds.) CIARP (1). Lecture Notes in Computer Science, vol. 8258, pp. 149–156. Springer (2013)
9. Fleuret, F.: Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research **5**, 1531–1555 (2004), http://www.jmlr.org/papers/volume5/fleuret04a/fleuret04a.pdf
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3**, 1157–1182 (2003)
11. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence **22**(1), 4–37 (2000)
12. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence **97**, 273–324 (1997)
13. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. on Knowledge and Data Engineering **17**(4), 491–502 (2005)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Peng, H., Long, F., Ding, C.H.Q.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell **27**(8), 1226–1238 (2005), http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.159
16. Rényi, A.: On Measures of Entropy and Information. In: Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability. pp. 547–561 (1961)
17. Smith, G.: On the foundations of quantitative information flow. In: Proc. of FOSSACS. LNCS, vol. 5504, pp. 288–302. Springer (2009)
18. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. CoRR **abs/1206.2944** (2012), http://arxiv.org/abs/1206.2944
19. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Computing and Applications **24**(1), 175–186 (2014)
20. Yang, H.H., Moody, J.: Feature selection based on joint mutual information. In: In Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis. pp. 22–25 (1999)