Neurocomputing 112 (2013) 64–78

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification

Benoît Frénay *, Gauthier Doquire [1], Michel Verleysen

Machine Learning Group—ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Available online 7 March 2013

Keywords:
Mutual information
Feature selection
Classification
Probability of misclassification
Hellman–Raviv and Fano bounds

ABSTRACT

Mutual information is a widely used performance criterion for filter feature selection. However, despite its popularity and its appealing properties, mutual information is not always the most appropriate criterion. Indeed, contrary to what is sometimes hypothesized in the literature, looking for a feature subset maximizing the mutual information does not always guarantee to decrease the misclassification probability, which is often the objective one is interested in. The first objective of this paper is thus to clearly illustrate this potential inadequacy and to emphasize the fact that the mutual information remains a heuristic, coming with no guarantee in terms of classification accuracy. Through extensive experiments, a deeper analysis of the cases for which the mutual information is not a suitable criterion is then conducted. This analysis allows us to confirm the general interest of the mutual information for feature selection. It also helps us better apprehending the behaviour of mutual information throughout a feature selection process and consequently making a better use of it as a feature selection criterion.

1. Introduction

Feature selection is known to be a preprocessing technique of fundamental importance for many applications in machine learning, pattern recognition or data mining. Indeed, dealing with high-dimensional data is a particularly hard task, in practice, due to many problems and counter-intuitive phenomena such as the empty space phenomenon and the concentration of distances [1,2]. Reducing the dimensionality of the datasets to a relatively low number of features is thus often necessary if one wants to build, for instance, efficient classification models. While efficient projection techniques can be used for dimensionality reduction, feature selection has the advantage of preserving the original features, which makes it possible to build easily interpretable models. Such an interpretability is highly appreciated, for instance, in the industrial and medical areas.

Among the various approaches to feature selection, filter methods are very popular and often used in practice. Filter methods are based on a relevance criterion independent of any classification model. They are thus easy to use and generally exhibit a low computational cost, especially when compared with wrapper methods which try to directly maximize the performances of a given prediction model. Filter methods have the additional advantage of being more general than wrapper or embedded methods, which perform simultaneously feature selection and prediction, in the sense that filters can be used in combination with any prediction model. The reader interested in feature selection is referred to [3] for a nice overview of this topic.

Since the seminal work of Battiti [4], the mutual information [5] has become one of the most widely used criteria for feature selection; see for example the following works [6–8]. In [6,7], the authors try to determine a set of maximally informative features which are mutually as non-redundant as possible, using the maximum relevance minimum redundancy principle and the conditional mutual information, respectively. In [8], a forward/backward search procedure is used to find the most relevant variables in spectroscopic modelling.

Besides performing often well in practice, the mutual information possesses other properties, detailed later in the paper, making it particularly well-suited for the feature selection task. These properties include the existence of bounds relating the mutual information to the probability of classification error. However, for a certain number of classification problems, mutual information is not the most appropriate choice of relevance criterion. Indeed, despite what is sometimes hypothesized, choosing a subset of features maximising the mutual information is not always equivalent to choosing a subset of features minimizing the

* Corresponding author. Tel.: +32 10 47 81 33; fax: +32 10 47 25 98.
E-mail addresses: benoit.frenay@uclouvain.be (B. Frénay), gauthier.doquire@uclouvain.be (G. Doquire), michel.verleysen@uclouvain.be (M. Verleysen).

[1] The author is funded by a Belgian F.R.I.A. Grant.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.051

misclassification probability, which is generally the quantity one is eventually interested in.

The first objective of this paper is thus to clearly point out this fact, by illustrating it through an intuitive example. Moreover, this work also aims at characterizing the problems for which the mutual information criterion is likely to fail, and in this case to which extend the loss in misclassification probability is important. To this end, extensive experiments have been carried out on both continuous and categorical datasets, either artificially generated or corresponding to real-world problems. The idea is to eventually assess the potential interest of the mutual information as a feature selection criterion, despite its non-optimality regarding the misclassification probability. This work extends preliminary results presented in [9]. Balanced datasets are considered and new experiments are conducted to gain a better insight on the behaviour of mutual information. A forward feature selection procedure is also analysed while only pairwise comparisons between features were considered in [9].

The rest of the paper is organised as follows. Section 2 briefly recalls basic definitions about the mutual information and details some of the reasons of its popularity for feature selection. Section 3 discusses and illustrates the potential inadequacy of the mutual information for feature selection; a problem for which mutual information is not appropriate is presented and a simple sufficient condition for its optimality is given. Sections 4–6 present the experimental results for artificial datasets with discrete features, artificial datasets with continuous features and real-world datasets with continuous features, respectively. Section 7 summarises the observations drawn from the experiments and Section 8 concludes the work.

2. Mutual information

The aim of this section is to remind fundamental notions about mutual information and to justify its interest for feature selection in classification problems.

2.1. Formal definitions

Shannon's mutual information [10,5] is a measure of the dependency existing between two random variables X and Y, considered to be discrete in this section. Let us assume that X (resp. Y) can take n_X (n_Y) possible different values x_i (y_i), each with probability $P_X(X=x_i)$ ($P_Y(Y=y_i)$). The mutual information is then defined as

$$I(X;Y)=\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}P_{XY}(X=x_i,Y=y_j)\times\log_2\frac{P_{XY}(X=x_i,Y=y_j)}{P_X(X=x_i)P_Y(Y=y_j)} \tag{1}$$

where $P_{XY}(X,Y)$ is the joint probability of the X and Y variables. Eq. (1) actually defines the Kullback–Leibler divergence [5] between the product of the two distributions $P_X(X)\times P_Y(Y)$ and the joint probability $P_{XY}(X,Y)$. As can be deducted from Eq. (1), the mutual information is a symmetric criterion, i.e. $I(X;Y)=I(Y;X)$.

Since the entropy of a discrete random variable X is defined as

$$H(X)=-\sum_{i=1}^{n_x}P_X(X=x_i)\log_2 P_X(X=x_i), \tag{2}$$

it can be shown [5] from Eq. (1) that the mutual information can be equivalently rewritten as

$$I(X;Y)=H(Y)-H(Y|X) \tag{3}$$

with

$$H(Y|X)=\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}P_{XY}(X=x_i,Y=y_j)\times\log_2\frac{P_X(X=x_i)}{P_{XY}(X=x_i,Y=Y_j)} \tag{4}$$

being the conditional entropy of Y once X is given. While the developments have been presented for discrete variables, similar definitions can as well be derived for continuous random variables. In this case, the sums are then replaced by integrals.

2.2. Interest for feature selection

Since the work of Battiti [4], the mutual information criterion has been used extensively for filter feature selection because of many desirable properties it possesses for this task.

The first important property of mutual information, as detailed in [4], is its natural interpretation in terms of uncertainty reduction. Indeed, the entropy is a measure of the uncertainty on the values taken by a random variable. Consequently, if Y denotes a target class vector and X is a (set of) feature(s), Eq. (3) shows that $I(X;Y)$ can be interpreted as the reduction of uncertainty about the value of Y once X is known. In this regard, mutual information is thus a quite intuitive criterion to maximize for a feature subset to be considered as good. If there is no dependency between X and Y, then $H(Y)=H(Y|X)$ and $I(X;Y)=0$. Similarly in Eq. (1), if X and Y are independent, $P_{XY}(X,Y)=P_X(X)P_Y(Y)$ and again $I(X;Y)=0$. On the contrary, if $Y=f(X)$, then the mutual information is maximal and $I(X;Y)=H(Y)$.

The second main advantage of the mutual information, as also stressed in [4], is that it is able to measure non-linear relationships between variables. Other criteria, such as the correlation coefficient, are limited to the detection of linear dependencies. The ability to detect non-linear dependencies is obviously a strong advantage since many of the most popular classification algorithms, such as support vector machines (with a non-linear kernel) and k-nearest-neighbors, are effectively able to model non-linear relationships between the features and the class label.

In addition, the mutual information criterion can naturally be defined for multivariate random variables (and thus for subsets of features), which is not true e.g. for the correlation coefficient. This is a property of major importance since greedy search procedures (such as forward, backward and forward/backward) are often used in practice to build a feature subset. This is because, in some situations, some features are only relevant or redundant when considered together. For example, in the well-known XOR problem, both features individually do not contain any information about the output, but together completely determine it. For such problem, a univariate criterion will never be able to detect any of the two features as relevant.

Finally, the use of mutual information for feature selection in classification problems is supported by the existence of bounds relating the misclassification probability P_e for an optimal classifier that achieves the Bayes risk to the conditional entropy $H(Y|X)$, where Y is again the class label and X the feature subset. Firstly, Fano [11] derived two lower bounds on P_e. The weaker bound is

$$H(Y|X)\le 1+P_e\log_2(n_Y-1) \tag{5}$$

where n_Y is the number of possible classes. The stronger bound states that

$$H(Y|X)\le H(P_e)+P_e\log_2(n_Y-1), \tag{6}$$

where $H(P_e)=-P_e\log_2 P_e-(1-P_e)\log_2(1-P_e)$ [5].

Both Eqs. (5) and (6) give bounds on $H(Y|X)$, but they can be inverted to provide a lower bound on P_e. However, the stronger bound is less easy to manipulate in practice than the weaker bound because P_e cannot be isolated in Eq. (5) in a closed form, the bound on P_e has thus to be computed numerically. Nevertheless, the stronger Fano bound in practice is much more useful than the weak one. Indeed, the weak bound (5) does not apply to binary classification problems, since in this case, Eq. (5) trivially reduces to $H(Y|X) \le 1$. Eq. (5) cannot thus be inverted to get a lower bound on P_e when $n_Y=2$. Moreover, the weak bound is generally much looser than the strong one. This is particularly true when P_e is small, which is however precisely the situation of interest for classifier design [12].

The probability of misclassification P_e can also be upper-bounded by the Hellman–Raviv inequality [13]

$$P_e \le \tfrac{1}{2}H(Y|X). \tag{7}$$

Fig. 1, inspired from [12,14], illustrates the three bounds introduced in Eqs. (5)–(7). As can be seen, decreasing the conditional entropy $H(Y|X)$ obviously decreases both the upper and the lower bound on P_e, which motivates the use of this criterion for feature selection. Notice that $H(Y)$ is a constant value for a given classification problem since it depends only on the class labels and not on the selected features. According to Eq. (3), maximizing the mutual information $I(X;Y)$ is thus equivalent to minimizing the conditional entropy $H(Y|X)$ in this context. Eqs. (5)–(7) give a justification to the maximisation of the mutual information for feature selection.

Notice that the upper bound (7) on P_e is an increasing concave, since it is linear with respect to $H(Y|X)$. Also, the lower bound (6) on P_e (as well as its weak form (5) which is linear with respect to $H(Y|X)$) is an increasing convex, since the converse upper bound on $H(Y|X)$

$$H(P_e)+P_e \log_2(n_Y-1) \ge H(Y|X) \tag{8}$$

is increasing concave with respect to P_e. Indeed, it can easily be shown that its first-order derivative

$$-\log_2 P_e+\log_2(1-P_e)+\log_2(n_Y-1) \tag{9}$$

is positive and that its second-order derivative

$$\frac{-\log_2 e}{P_e(1-P_e)} \tag{10}$$

is negative when $P_e \le ((n_Y-1)/n_Y)$, which is the case since P_e is the misclassification probability for an optimal classifier. These properties are used in the next section.

3. Potential inadequacy of mutual information

As mentioned in Section 1, the actual objective of feature selection is often to reduce as much as possible the probability of misclassification of a model built on the selected feature subset. In other words, the quality and the utility of a feature subset can be measured through P_e, which actually gives a lower bound for the misclassification probability of any (suboptimal) classification model. Based on the convex lower bound and the concave upper bound in Fig. 1 and Eq. (3), several papers, e.g. [12,14,15], claim that a feature subset having a higher mutual information with the output than another one will lead to a smaller probability of misclassification P_e. Those papers conclude that the mutual information can therefore be used as a proxy for P_e in a feature selection context. The objective of this section is to show that such a conclusion is not always valid in practice. A simple condition for the optimality of the mutual information as a feature selection criterion is also given. Eventually, we also derive a bound relating (i) the maximum value of mutual information between two feature subsets and the output to (ii) the loss in misclassification probability induced by the selection of one subset instead of the other one.

3.1. Relationship between misclassification probability and conditional entropy

In Fig. 2, the strong Fano bound and the Hellman–Raviv bound for P_e in terms of $H(Y|X)$ are again illustrated. Moreover, the figure also shows many examples of $\langle H(Y|X),P_e \rangle$ couples of values. Each point in Fig. 2 corresponds to a different random binary classification problem with two binary features. The problems are generated as follows: (i) the two values $P(Y=y)$ (for $y \in [0,1]$) and the four values $P(X=x|Y=y)$ (for $x \in [0,1]$ and $y \in [0,1]$) are randomly drawn from the uniform distribution $\mathcal{U}(0,1)$, (ii) these values are normalised to ensure that they represent probabilities, i.e. $\sum_y P(Y=y)=1$ and $\sum_x P(X=x|Y=y)=1$ for each y and (iii) probabilities $P(X)$ and $P(Y|X)$ are eventually computed using marginalisation and the Bayes' theorem. For each problem, it is thus possible to compute exactly both P_e and $H(Y|X)$ because all the necessary probabilities are known.

As expected, the couples $\langle H(Y|X),P_e \rangle$ all lie in the area defined by the strong Fano lower bound and the Hellman–Raviv upper bound. Given the value of the mutual information $I(X;Y)$, or equivalently the value of the conditional entropy $H(Y|X)$, the two bounds thus define an interval where P_e belongs. Obviously, given two different values of conditional entropy, the intervals for

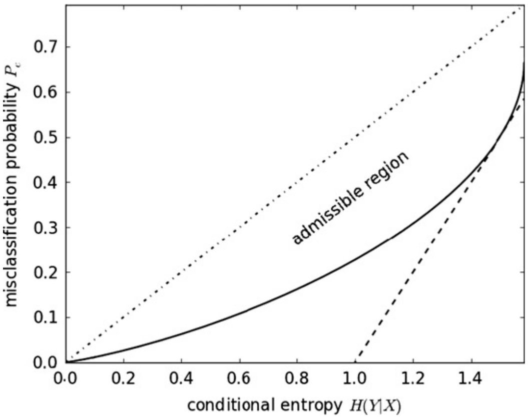

Fig. 1. Weak Fano bound (dashed line), strong Fano bound (plain line) and Hellman–Raviv bound (dash-dotted line) on the misclassification probability P_e of an optimal classifier with three classes (n_Y=3), in terms of the conditional entropy $H(Y|X)$; figure inspired from [12,14], reprinted with permission from [9].

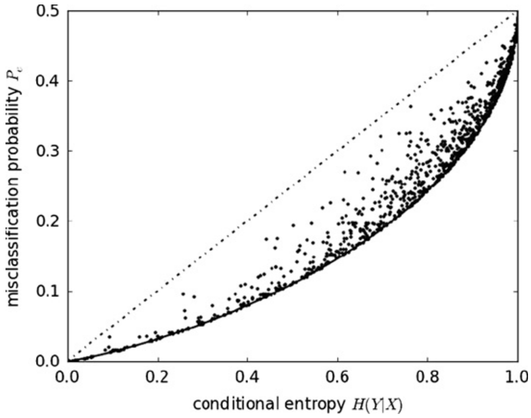

Fig. 2. Several pairs of $\langle H(Y|X),P_e \rangle$ values corresponding to random binary classification problems with two binary features. The strong Fano bound (plain line) and the Hellman–Raviv bound (dash-dotted line) relating P_e to $H(Y|X)$ are shown. From [9], reprinted with permission.

P_e defined by the bounds could strongly overlap. Therefore, given two subsets of features $\mathcal{X}_1$ and $\mathcal{X}_2$ such that $H(Y|\mathcal{X}_1) < H(Y|\mathcal{X}_2)$, it could theoretically be possible that $\mathcal{X}_1$ leads to a higher probability of misclassification P_e than $\mathcal{X}_2$. Fig. 2 illustrates the fact that, in practice, this situation could actually happen. Indeed, even if the pairs $\langle H(Y|X), P_e \rangle$ mainly lie near the lower bound, they scatter the whole area between the two bounds; for a given conditional entropy, actual values of misclassification probability can thus be obtained in the whole interval defined by the bounds. Consequently, choosing between two feature sets based on the mutual information criterion could not be optimal (in terms of misclassification probability), as shown through a simple example in Section 3.2.

3.2. Illustration of mutual information failure for feature selection

A simple example is now presented, to illustrate the potential inadequacy of the mutual information in a feature selection context. Let us consider a disease diagnosis, where two classes have the same prior probability

$$P(Y) = (0.5 \;\; 0.5). \tag{11}$$

In (11), each column corresponds to one of the two possible values of $Y \in \{0,1\}$. Let us further assume that the results of two different tests are available to help classifying a new patient. Both tests are binary and their outcomes are denoted as $X_1 \in \{0,1\}$ and $X_2 \in \{0,1\}$. For some practical reasons, the practician can only perform one of those two tests. This choice is clearly a feature selection problem, each test corresponding to a feature and the practician having to chose the best test.

Through previous experimentation, the practician is able to establish that the conditional distributions $P(X_i|Y)$ of both tests X_1 and X_2 given Y are given by

	$Y=0$	$Y=1$
$X_1=0$	0.287	0.758
$X_1=1$	0.713	0.242

and

	$Y=0$	$Y=1$
$X_2=0$	0.627	0.999.
$X_2=1$	0.373	0.001

The rows correspond to the possible values of X_i and the columns again correspond to the values of Y. Using marginalisation and the Bayes' theorem, it is straightforward to obtain the posteriors $P(Y|X_i)$ given by

	$X_1=0$	$X_1=1$
$Y=0$	0.275	0.746
$Y=1$	0.725	0.254

and

	$X_2=0$	$X_2=1$
$Y=0$	0.385	0.999.
$Y=1$	0.615	0.001

Again, rows correspond to values of Y and columns correspond to values of X_i. It can be understood from the last two probability tables that the test whose outcome is X_1 allows discriminating fairly well between the classes, whatever its output is. There remains however a quite important misclassification probability using X_1 (P_e=0.275 if $X_1 = 0$ and P_e=0.254 if $X_1 = 1$). The second test, with the outcome X_2, allows discriminating almost perfectly when it is positive (P_e=0.001 if $X_2 = 1$). When it is negative ($X_2 = 0$), it is however much less discriminative than the first test since the misclassification probability is P_e=0.385 in that case.

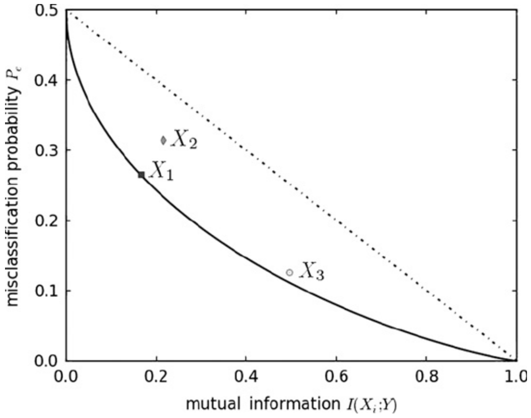

Fig. 3. Example of mutual information failure for feature selection, with the strong Fano bound (plain line) and the Hellman–Raviv bound (dash-dotted line).

When the results of the first tests are used to select the feature, one eventually obtains a global misclassification probability of P_e=0.265 while $I(X_1;Y) = 0.167$. Using the second test, one obtains P_e=0.314 and $I(X_2;Y) = 0.217$. Here, it appears that the mutual information is larger using X_2. However, P_e is smaller when X_1 is used, meaning that selecting X_2 based on mutual information leads here to an increased probability of misclassification.

The above example is illustrated in Fig. 3, where each point $\langle H(Y|X), P_e \rangle$ is again shown to lie between the Fano and Hellman–Raviv bounds. Obviously, $I(X_2;Y) = H(Y) - H(Y|X_2)$ is larger than $I(X_1;Y) = H(Y) - H(Y|X_1)$ while $P_e(X_2)$ is simultaneously larger than $P_e(X_1)$.

3.3. A condition of optimality

As illustrated by the previous example, and as shown in the following sections, the mutual information appears to be a heuristic with no obvious way to assess its potential interest. However, in some situations, it is possible to guarantee that the mutual information is actually an adequate criterion. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two features sets that have to be compared. If the value of the Hellman–Raviv bound for $\mathcal{X}_1$ is smaller than the value of the strong Fano bound for $\mathcal{X}_2$, then it can be deduced from the bounds in Fig. 2 that the feature set $\mathcal{X}_1$ leads to a smaller misclassification probability P_e than the feature set $\mathcal{X}_2$ does. Thus, if the values of the conditional entropies for two subsets are different enough, the corresponding possible intervals for P_e cannot overlap and ranking feature subsets with the mutual information criterion is optimal.

In the above example, the Fano bound for X_1 is $P_e \geq 0.264$, whereas the Hellman–Raviv bound for X_2 is $P_e \leq 0.391$; it is not possible to guarantee that mutual information is a relevant criterion to choose between X_1 and X_2. Fig. 3 also shows another candidate X_3 for which the Hellman–Raviv bound is $P_e \leq 0.252$. In this case, the new feature X_3 is guaranteed to be a better choice.

3.4. Upper bound on the misclassification probability loss

It is also possible to give an upper bound for the difference in misclassification probability in case of failure, i.e. the supplementary percentage of samples which are misclassified due to an incorrect choice of feature subset only. This difference is called the misclassification probability loss in the following of the paper. Indeed, the worst case of mutual information failure occurs when (i) both feature subsets have almost identical mutual information, (ii) the selected feature subset stands on the Hellman–Raviv bound (maximum misclassification probability) and (iii) the other feature subset stands on the strong Fano bound (minimum misclassification probability). In such a case, the misclassification probability loss is simply the difference between the Hellman–Raviv bound and the strong Fano bound. The upper bound on the misclassification probability loss is concave with respect to $I(X;Y)$, since the Hellman–Raviv and Fano bounds are increasing concave and convex with respect to $H(Y|X)$, respectively. Fig. 4 shows the upper bound on the misclassification probability loss for the above example. Here, the misclassification probability loss is bounded by 0.159 for the selected feature X_2, whereas the actual misclassification probability loss is 0.049. Interestingly, the maximum misclassification probability loss decreases for extreme (small or large) values of the mutual information. It suggests that mutual information failures have less important consequences in these cases.

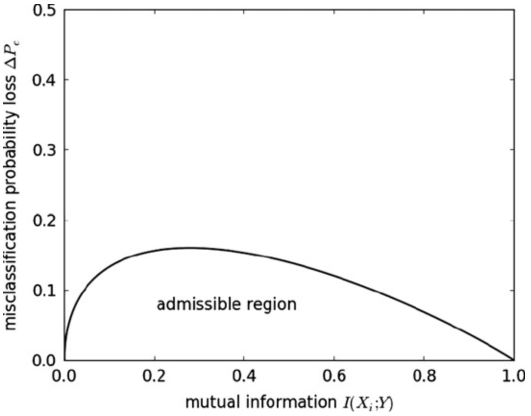

Fig. 4. Theoretical upper bound on the misclassification probability loss for binary classification with balanced classes.

4. Artificial classification problems with discrete features

This section discusses the use of mutual information for feature selection using three simple artificial monovariate binary classification problems. The input of the classifier is a discrete feature with 2^d possible modalities. This may be viewed as equivalent to a binary classification problem with d binary features.

4.1. Experimental settings

The three artificial problems discussed in this section are designed to simulate low, medium and high levels of difficulty in binary classification. This is achieved by choosing different

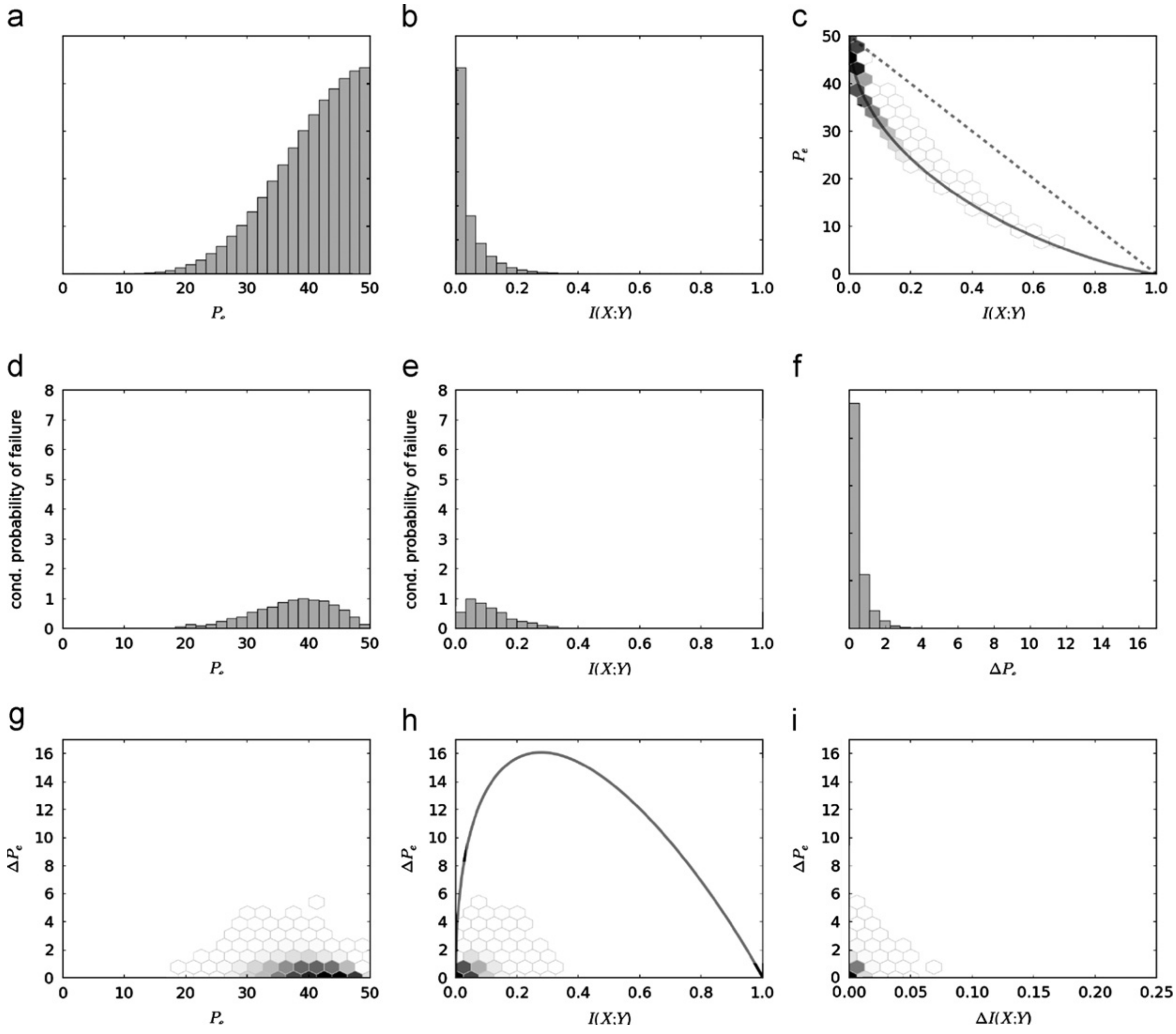

Fig. 5. Results for artificial binary classification problems with a 2-value discrete feature and a Dirichlet prior with $\alpha = 4$ for the conditional probabilities $P(X = x_i|Y)$.

domain sizes for the discrete feature X and different prior distributions for its conditional probabilities $P(X|Y)$. For each of the three problems, a large number of pairs of possible features are generated, which are compared pairwise to assess whether mutual information is consistent with the misclassification probability. Notice that the classifier remains univariate: the possible features are compared by pairs, to decide in each pair which feature will be used as input to the classifier. The conditional probabilities $P(X=x_i|Y)$ of each feature are drawn from a symmetric Dirichlet distribution

$$f(x_1,\dots,x_m|\alpha)=\Gamma(\alpha m)\prod_{i=1}^{m}\frac{x_i^{\alpha-1}}{\Gamma(\alpha)} \tag{12}$$

where x_i is the ith modality of feature X, Γ is the gamma function and α is the concentration parameter. Conditional probabilities $P(X|Y)$ are drawn instead of conditional probabilities $P(Y|X)$ because it allows us to keep constant the class prior $P(Y)$. Indeed, the proportion of instances in each class should not depend on the feature which is used to classify them. Moreover, this is necessary to compute the bounds which are visualised in the figures below. Large values of α correspond to conditional distributions of X given Y where almost all probabilities are equal, whereas only one probability is non-zero for small values of α. In other words, class discrimination is expected to be easier with small values of α. In addition, classes are usually easier to discriminate in high-dimensional spaces. By choosing the problem parameters $\langle m=2,\alpha=4\rangle$, $\langle m=8,\alpha=1\rangle$ and $\langle m=128,\alpha=0.06\rangle$, three families of problems are obtained with high, medium and low levels of difficulty, respectively. The class prior is uniform, i.e. $P(Y=y)=\frac{1}{2}$ for each y, in order to avoid class imbalance effects.

For each set of binary classification problem parameters, 10^6 pairs of features are generated. For each pair, the features are compared in terms of mutual information with the class Y and misclassification probability, which can be computed exactly since all the required probabilities are known. The feature with the largest mutual information is chosen. If the misclassification probability is also larger for the chosen feature, the pair is an example of failure for mutual information as a feature selection criterion. In case of failure, the difference in misclassification probability is called the misclassification probability loss ΔP_e, i.e. the percentage of samples which are misclassified due to an incorrect choice of feature. The average and conditional probabilities of failure can be estimated by counting failures among the pairs.

Each artificial problem is illustrated in Figs. 5–7, respectively. Mutual information is computed in base 2, whereas all probabilities are given in percents. The first row consists of a histogram of

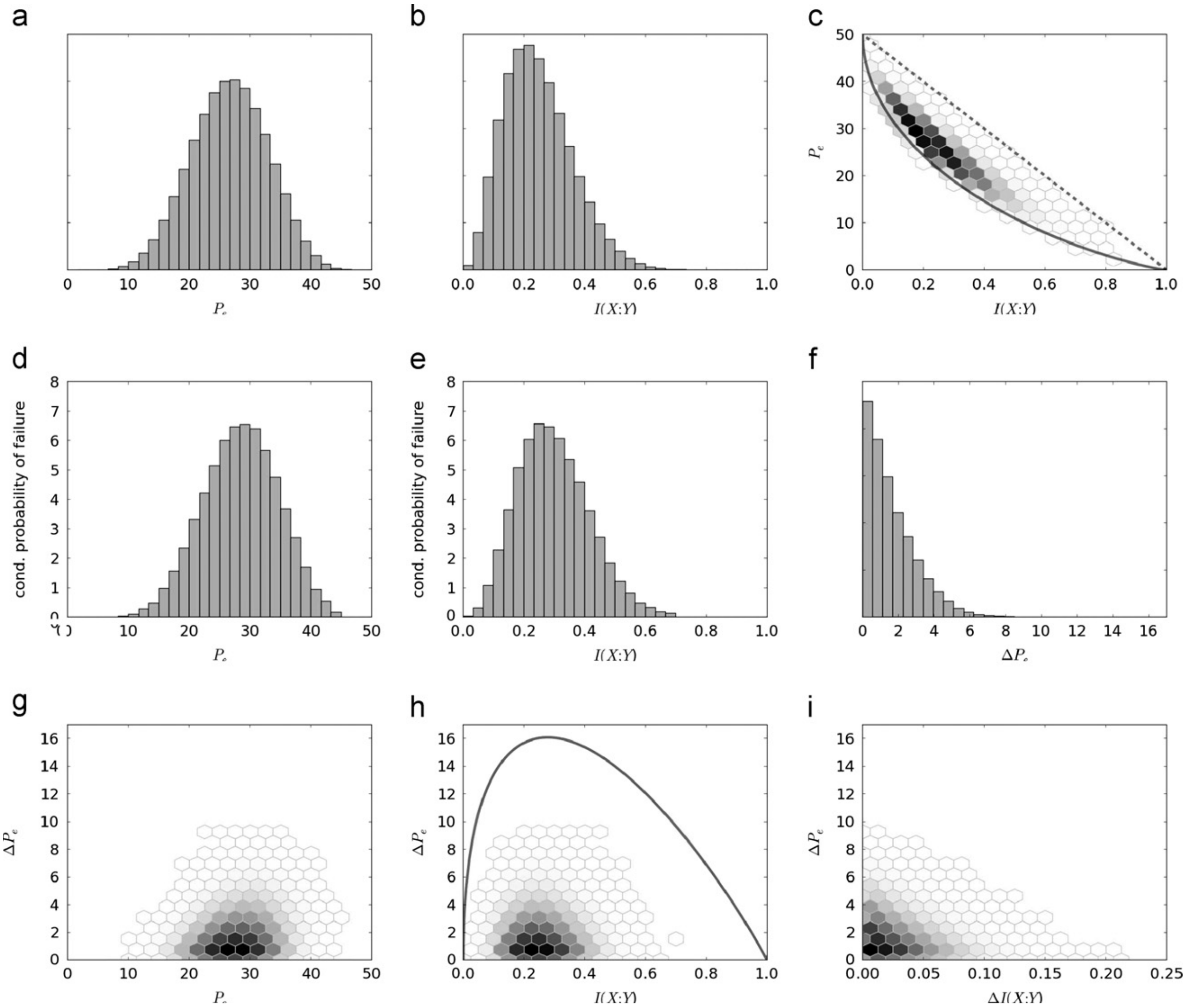

Fig. 6. Results for artificial binary classification problems with a 8-value discrete feature and a Dirichlet prior with $\alpha=1$ for the conditional probabilities $P(X=x_i|Y)$.

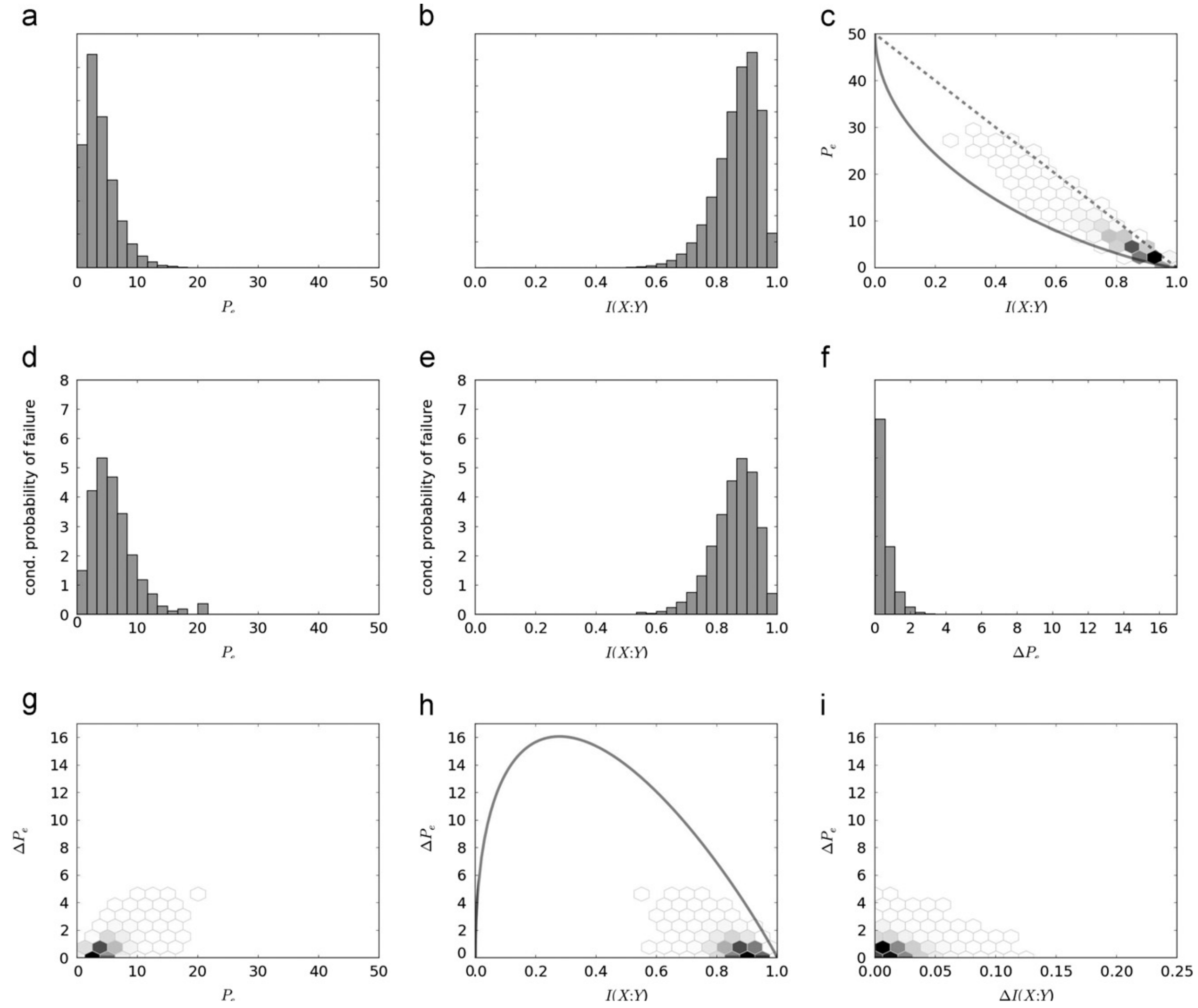

Fig. 7. Results for artificial binary classification problems with a 128-value discrete feature and a Dirichlet prior with $\alpha=0.06$ for the conditional probabilities $P(X=x_i|Y)$.

the misclassification probability, a histogram of the mutual information and a two-dimensional histogram (with hexagonal bins whose opacity indicates the number of samples in each bin) of these two quantities, with the Fano and Hellman–Raviv bounds. The second row shows an estimate of the conditional probability of failure given the misclassification probability and given the mutual information, and a histogram of the misclassification probability loss in case of failure. Eventually, for failures, the last row shows two-dimensional histograms of (i) the misclassification probability and the misclassification probability loss, (ii) the mutual information and the misclassification probability loss (with the theoretical bound derived in Section 3.4) and (iii) the mutual information difference and the misclassification probability loss. In the last two rows, which correspond to mutual information failures, the mutual information and the misclassification probability are those of the feature which is selected using mutual information. Indeed, what we are mainly interested in is to know when a feature selected by mutual information is likely to be a bad choice.

4.2. Results

For $m=2$ and $\alpha=4$, classes are very difficult to discriminate, as seen in Fig. 5(a) and (b), but only 0.6% of the pairs are failures. The conditional probability of failure remains small in Fig. 5(d) and (e), where the failure probability decreases for small mutual information values and large misclassification probabilities. Fig. 5(f) shows that 95% of the misclassification probability losses remain below 1.5%. In Fig. 5(g) and (h), the misclassification probability loss decreases for small mutual information values and large misclassification probabilities. Eventually, Fig. 5(i) shows that failures occur when comparing pairs of features which are close in terms of both mutual information and misclassification probability.

For $m=8$ and $\alpha=1$, classes are moderately difficult to discriminate, as seen in Fig. 6(a) and (b). The percentage of failure is 4.6, what is higher than in the $m=2$ and $\alpha=4$ case. The conditional probability of failure is also larger in Fig. 6(d) and (e); Fig. 6(f) shows that the misclassification probability loss is larger, but remains below 4.3% in 95% of the failures. Fig. 6(g)–(i) leads to similar conclusions than with $m=2$ and $\alpha=4$.

For $m=128$ and $\alpha=0.06$, classes are quite easy to discriminate, as seen in Fig. 7(a) and (b). The percentage of failure is 3.8 and the conditional failure probability decreases for large mutual information values and small misclassification probabilities. Similarly, Figs. 7(g) and (h) show that the misclassification probability loss decreases for large mutual information values and small misclassification probabilities. In Fig. 7(f), 95% of the misclassification probability losses remain below 1.5%.

4.3. Discussion

In the above experiments, mutual information fails to select the feature with the best misclassification probability in only a few percents of the cases. Moreover, such failures do not lead to large misclassification probability losses, which means that the consequences of the failures are not too important. Failures appear to be more probable when classes are moderately difficult to discriminate, i.e. for intermediate values of mutual information and misclassification probability. In such cases, the misclassification probability loss is also larger. For all problems, failures mostly occur for pairs of features which are close in terms of both mutual information and misclassification probability.

5. Artificial classification problems with continuous features

This section discusses the use of mutual information for feature selection using three simple artificial three-class classification problems with a single continuous feature.

5.1. Experimental settings

Similar to Section 4, the three artificial problems discussed in this section are designed to simulate low, medium and high levels of difficulty in three-class classification. For each of the three problems, a large number of features are generated, which are compared pair-wise to assess whether mutual information is consistent with the misclassification probability. For each class, each feature has a unidimensional Gaussian conditional distribution. The standard deviations of the feature values are randomly drawn from a gamma distribution

$$f(\sigma|k,\theta)=\frac{\sigma^{k-1}}{\theta^k\Gamma(k)}e^{-\sigma/\theta}, \tag{13}$$

where k is the shape parameter and θ is the scale parameter. In the experiments, the parameter values $k=2$ and $\theta=0.5$ are used in order to obtain realistic and diversified standard deviations. The means of the feature values in the three classes are $\mu_0=-\Delta$, $\mu_1=0$ and $\mu_2=\Delta$, where Δ is a parameter which determines the difficulty of the classification problem. Large values of Δ correspond to easy problems with well-separated Gaussian distributions, whereas difficult problems with overlapping Gaussian distributions are obtained for small values of Δ. The problem parameters are $\Delta=0.5$, $\Delta=2$ and $\Delta=4$ and the class priors are uniform, i.e. $P(Y=y)=\frac{1}{3}$ for each y. Fig. 8 shows an example for each difficulty of three-class classification problem.

Similar to Section 4, 10^6 pairs of features are generated for each of the three-class classification problems. In each pair, the two features are compared in terms of mutual information and misclassification probability. Mutual information is computed in base 2, whereas all probabilities are given in percents. Since the distribution $P(X)$ is a mixture of Gaussian distributions, it is impossible to obtain exact values for the entropy $H(X)$ and the mutual information. Only the conditional entropy $H(Y|X)$ and the conditional probabilities $P(X|Y)$ can be computed analytically. In order to solve this problem, for each feature, 10^4 samples are drawn from each class. The conditional probabilities $P(X|Y)$ of the samples are computed analytically and used to obtain an estimate of the mutual information and the misclassification probability. Given the large number of samples and the low dimensionality of the data, the estimates are expected to be accurate. However, to deal with approximation errors, failures with a mutual information difference below 0.01 or a misclassification probability loss below 0.1% are ignored. Remaining computations and Figs. 9–11 are obtained similarly to Section 4.

5.2. Results

For $\Delta=0.5$, the three classes are quite difficult to discriminate, as seen in Fig. 9(a) and (b). The percentage of failures is 1.2 and the failure probability decreases for extreme (i.e. small or large) mutual information values and extreme misclassification probabilities in Fig. 9(d) and (e). Fig. 9(f) shows that 95% of the misclassification probability losses remain below 3.2%. In Fig. 9(g) and 9(h), the misclassification probability loss decreases for extreme mutual information values and misclassification probabilities. Eventually, Fig. 9(i) shows that failures mostly occur for pairs of features which are close in terms of both mutual information and misclassification probability.

For $\Delta=2$, classification is of medium difficulty, as seen in Fig. 10(a) and (b). The percentage of failure is 0.9 and the failure probability decreases for large mutual information values and small misclassification probabilities in Fig. 10(d) and (e). Fig. 10(f) shows that 95% of the misclassification probability losses remain below 2.7%. In Fig. 10(g) and (h), the misclassification probability loss decreases for large mutual information values and small misclassification probabilities. Again, failures occur for pairs of features which are close in terms of both mutual information and misclassification probability, as seen in Fig. 10(i).

For $\Delta=4$, the three classes are quite easy to discriminate, as seen in Fig. 11(a) and (b). The percentage of failure is 0.2 and the failure probability decreases for large mutual information values and small misclassification probabilities in Fig. 11(d) and (e). Fig. 11(f) shows that 95% of the misclassification probability losses remain below 1%. Figs. 11(g)–(i) lead to similar conclusions than for $\Delta=2$.

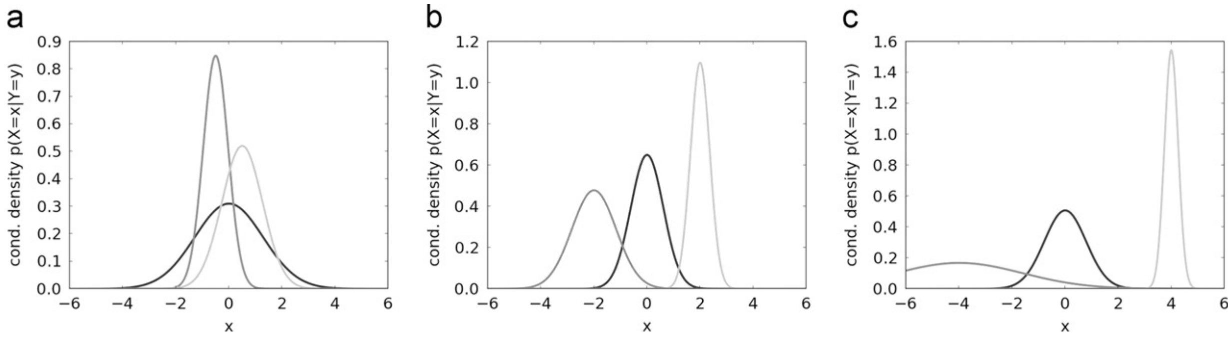

Fig. 8. Examples of three-class balanced classification problems of various difficulties. Standard deviations of the Gaussian distributions are randomly drawn from a gamma distribution with shape $k=2$ and scale $\theta=0.5$, whereas centers are chosen using $\Delta=0.5$, $\Delta=2$ and $\Delta=4$.

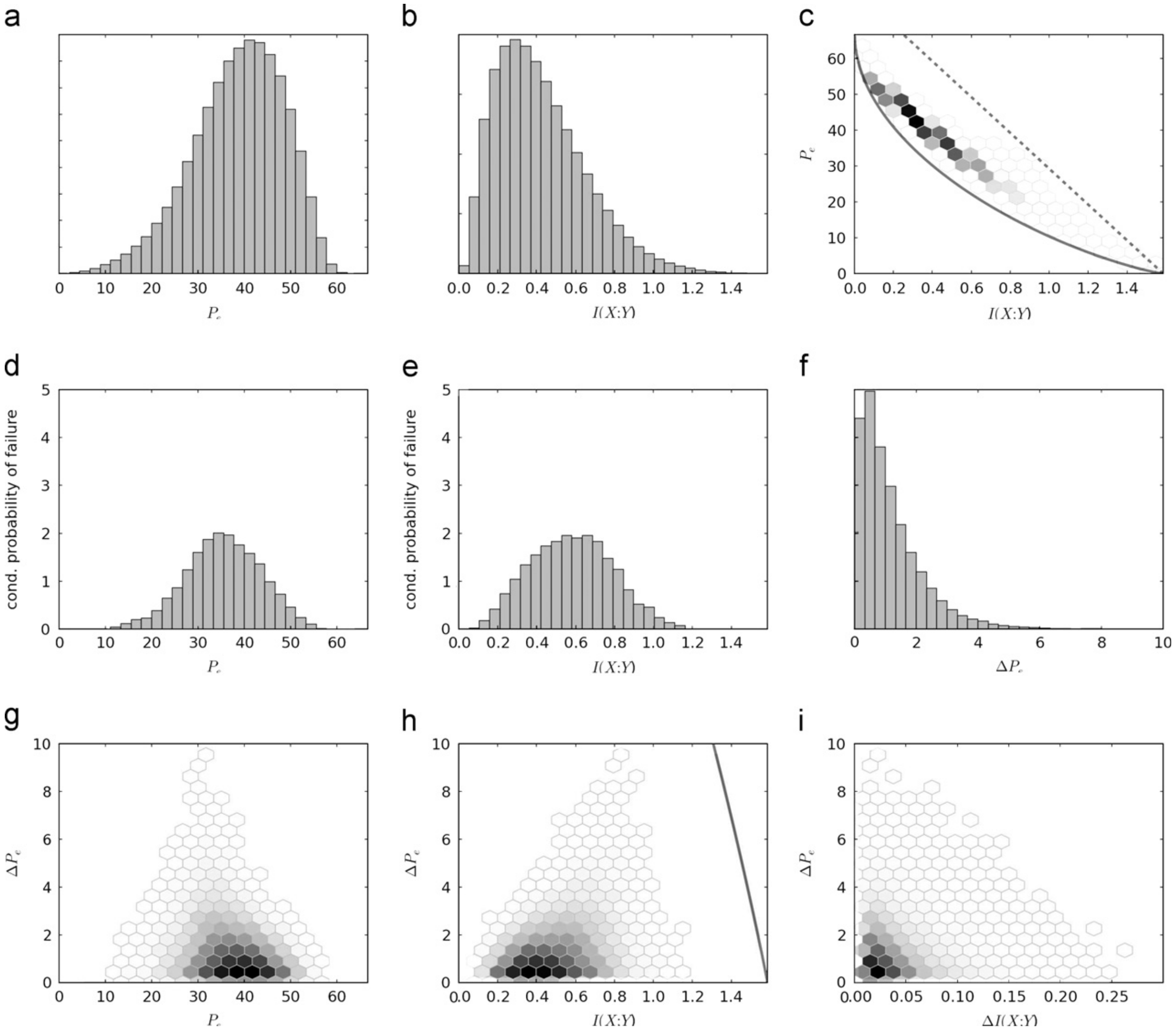

Fig. 9. Results for 10^6 pairs of artificial three-class problems with one continuous feature. Class distributions are Gaussians centered at $x = -0.5$, $x = 0$ and $x = 0.5$ and whose widths are randomly drawn from a gamma distribution. Mutual information values and misclassification probabilities are estimated using 10^4 samples from each class.

5.3. Discussion

The lessons of the above experiments are similar to those of the experiments in Section 4. Mutual information fails to select the feature with the best misclassification probability in only a few percents of the cases and the misclassification probability loss remains quite small. Again, failures appear to be more probable and to have more important consequences when classes are moderately difficult to discriminate, i.e. for intermediate values of mutual information and misclassification probability. For all problems, failures mostly occur for pairs of features which are close in terms of both mutual information and misclassification probability.

6. Real-world classification problems with continuous features

This section discusses the use of mutual information for feature selection using three real-world classification problems with continuous features. Feature selection is performed using a mutual information-based forward search algorithm [16], with the aim of assessing whether mutual information failures are more likely to occur at certain stages of a multivariate feature selection process.

6.1. Experimental settings

This section presents the results obtained with real-world datasets from the UCI repository [17]. Three balanced datasets with a large number of instances are chosen, in order to obtain reliable estimates of the mutual information and the misclassification probability. Firstly, Digits is a 10-class digit recognition dataset which contains 10,992 instances with 16 continuous features. Secondly, Wallrobot is a two-class robot navigation dataset which contains 4302 instances with 24 continuous features. The original dataset contains four classes, but only the two majority classes are kept in order to obtain a balanced dataset. Eventually, Wave is a three-class waveform dataset which contains 5000 instances with 21 continuous features. All datasets are almost perfectly balanced.

For each dataset, the feature selection process is repeated 5000 times. For each repetition, 10 features are randomly chosen among the set of available features in order to obtain a subproblem whose characteristics remain similar to the full problem. Then, a forward search is performed to find feature subsets of

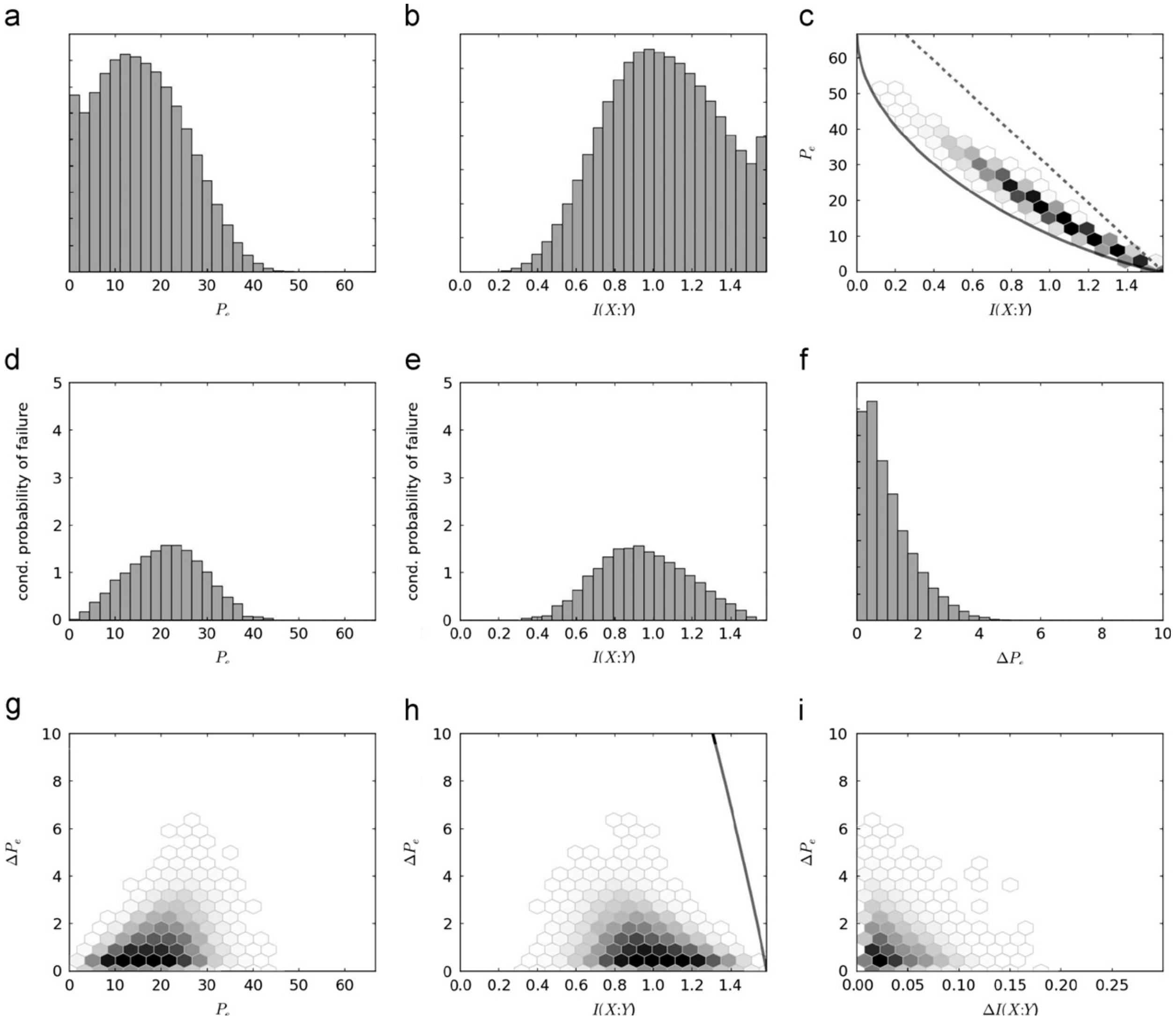

Fig. 10. Results for 10^6 pairs of artificial three-class problems with one continuous feature. Class distributions are Gaussians centered at $x=-2$, $x=0$ and $x=2$ and whose widths are randomly drawn from a gamma distribution. Mutual information values and misclassification probabilities are estimated using 10^4 samples from each class.

increasing sizes. The selection criterion is the mutual information, which is estimated as detailed below. For each forward step in each repetition, the misclassification probabilities are also estimated. A forward step is a failure if the feature subset which is selected in order to maximise the mutual information does not minimise the misclassification probability, i.e. if there exists a feature subset with a lower misclassification probability at this step. The mutual information and the misclassification probabilities are directly estimated using the conditional probabilities $P(Y|X)$ of each sample. These conditional probabilities are obtained from the conditional probabilities $P(X|Y)$, which are estimated using the Kozachenko–Leonenko estimator [18], by using the Bayes rule and marginalisation. Mutual information is computed in base 2, whereas all probabilities are given in percents. Moreover, in case of failure, the mutual information difference and the misclassification probability loss are computed between the feature with the best mutual information and the feature with the best misclassification probability. Fig. 12 shows a two-dimensional histogram of the mutual information and the misclassification probability for each dataset. The Fano and Hellman–Raviv bounds hold, what illustrates the validity of the above procedure.

Figs. 13–15 show several plots for each real-world problem. The first row consists of a histogram of the misclassification probability, a histogram of the mutual information and the misclassification probability for different feature subset sizes. The second row shows an estimate of the conditional probability of failure given the misclassification probability and given the mutual information, and the mutual information for different feature subset sizes. The third row shows two-dimensional histograms of (i) the misclassification probability and the misclassification probability loss, (ii) the mutual information and the misclassification probability loss (with the theoretical bound derived in Section 3.4) and (iii) the mutual information difference and the misclassification probability loss. The fourth row shows the mutual information difference and the misclassification probability loss for different feature subset sizes, and an estimate of the conditional probability of failure given the feature subset size. In the last three rows, which correspond to mutual information failures, the mutual information and the misclassification probability are those of the feature which is selected using mutual information. Indeed, what we are mainly interested in is to know when a feature selected by mutual information is likely to be a bad feature in terms of misclassification probability.

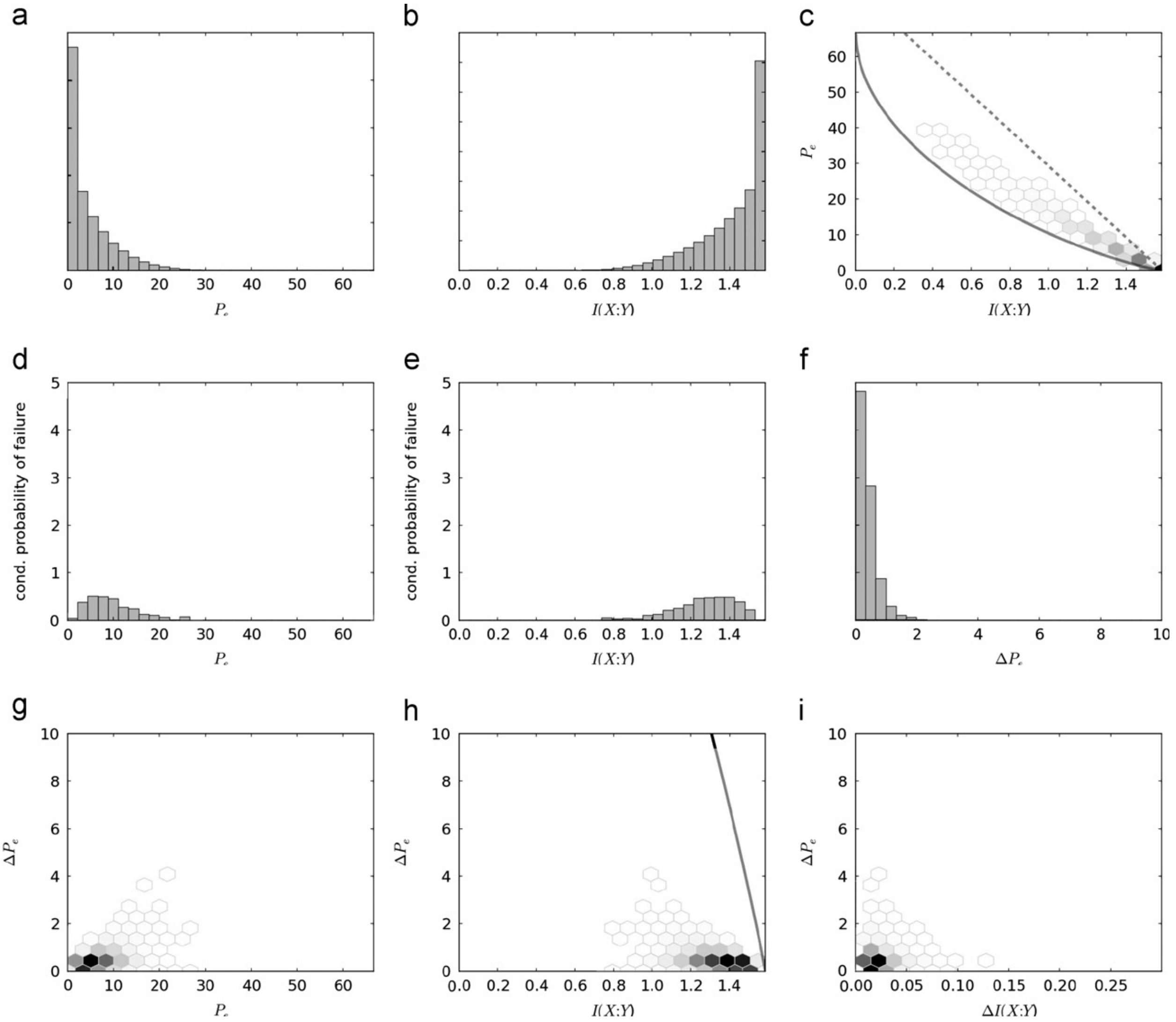

Fig. 11. Results for 10^6 pairs of artificial three-class problems with one continuous feature. Class distributions are Gaussians centered at $x=-4$, $x=0$ and $x=4$ and whose widths are randomly drawn from a gamma distribution. Mutual information values and misclassification probabilities are estimated using 10^4 samples from each class.

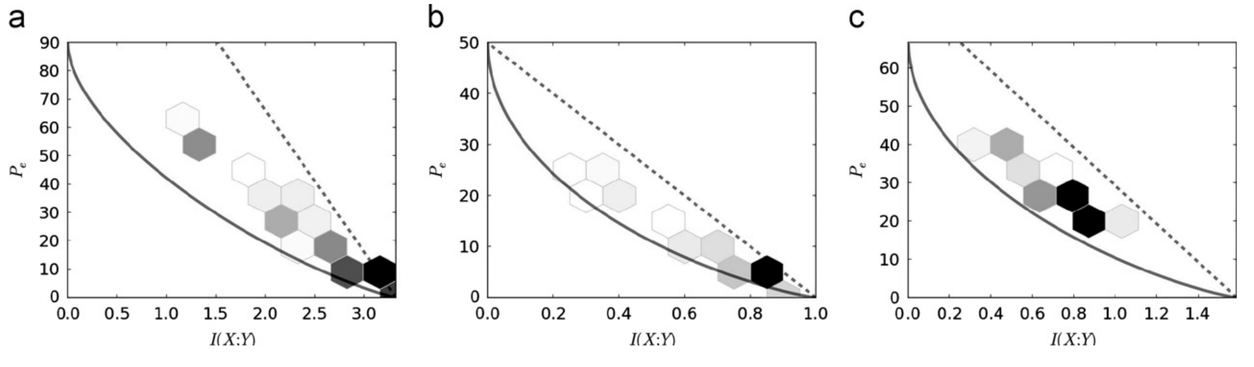

Fig. 12. Mutual information values and misclassification probabilities with random subsets of 10 features for the Digits, Wallrobot and Wave datasets, with Fano and Hellman–Raviv bounds.

6.2. Results

Figs. 13(a)–(c) and (f) show that forward search goes through a wide range of problem difficulties for the Digits dataset. As the feature subset size increases, the misclassification probability decreases slowly. The percentage of failures is 7.8, but Figs. 13(g)–(i) show that 95% of the misclassification probability loss remain below 2%. The dark hexagonal bin in these figures indicates the failures occur when the feature with the best mutual information and the feature with the best misclassification probability are close in terms of both mutual information and misclassification probability. The misclassification probability loss decreases for large mutual information values and small misclassification probabilities. In Fig. 13(d) and (e), the

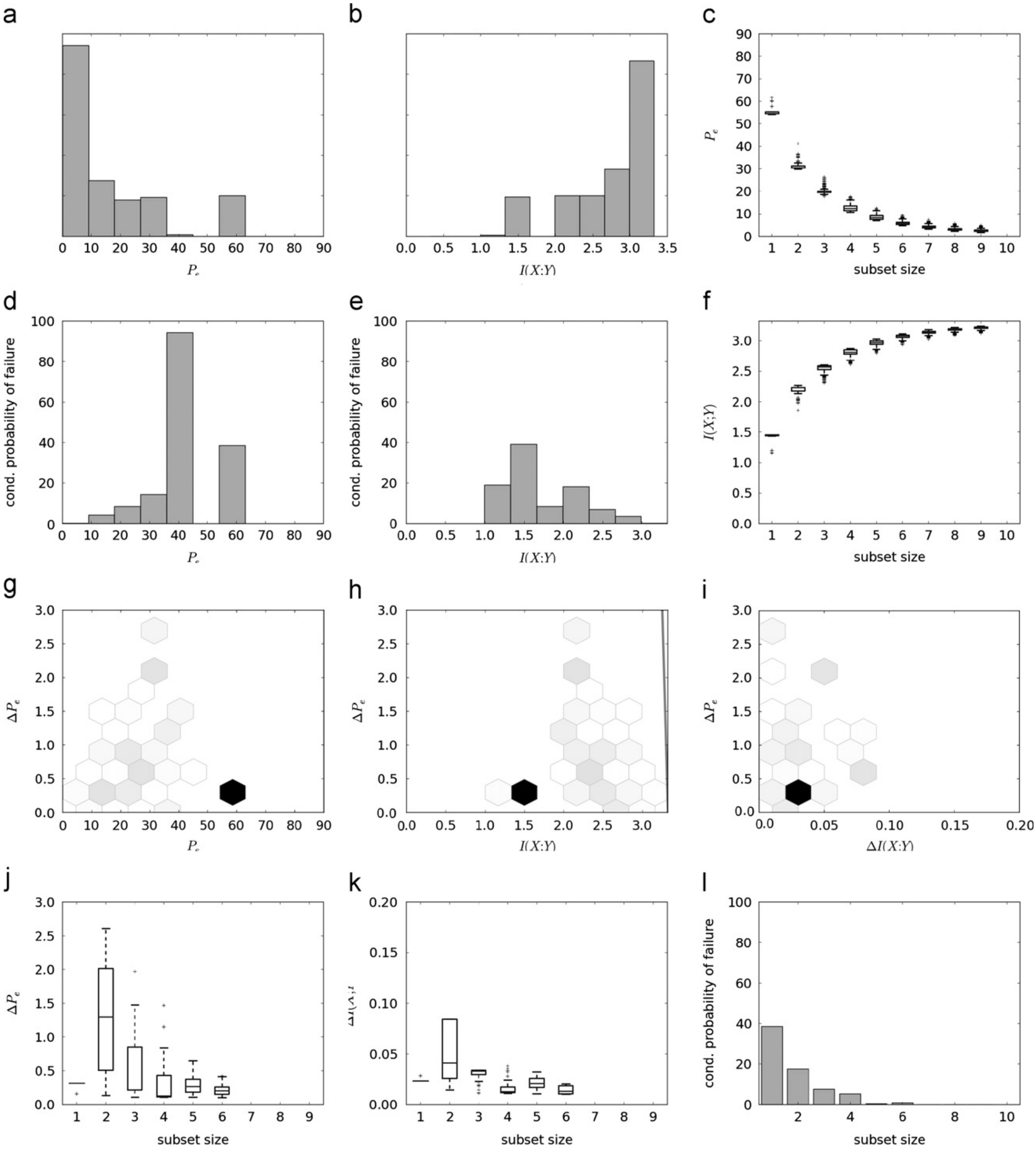

Fig. 13. Results of mutual information-based forward search with random subsets of 10 features for the Digits dataset.

misclassification probability loss decreases for large mutual information values and small misclassification probabilities. Figs. 13(j)–(l) show that the probability of failure is maximum at the beginning of the forward search, where the misclassification probability loss is small, and decreases quickly as the features subset size increases.

The results for the Wallrobot dataset are similar to the result for the Digits dataset, except (i) that classification performances are already optimal with about three features, as seen in Fig. 14(c) and (f), and (ii) that failures are much more likely to occur at the first step of the forward search, as seen in Fig. 14(l), what corresponds to intermediate values of mutual information and the misclassification probability. Consequently, Fig. 14(j) shows that the misclassification probability loss is only significant for subsets with a single feature. It corresponds to the peak of probability of failure in Fig. 14(d) and (e) and to the dark hexagonal bin in Fig. 14(g)–(i). The percentage of failures is 2.4 and 95% of the misclassification probability losses remain below 1%.

Figs. 15(a)–(c) and (f) show that the Wave dataset corresponds to a quite difficult problem. The percentage of failure is 0.2. Contrary to the Digits and Wallrobot datasets, the conditional probability of failure in Fig. 15(l) first increases for small features subset sizes, achieves its maximum for three features and then quickly decreases. The misclassification probability is large for the two first feature subset sizes in Fig. 15(c), what suggests again that failures more likely occur for intermediate mutual information

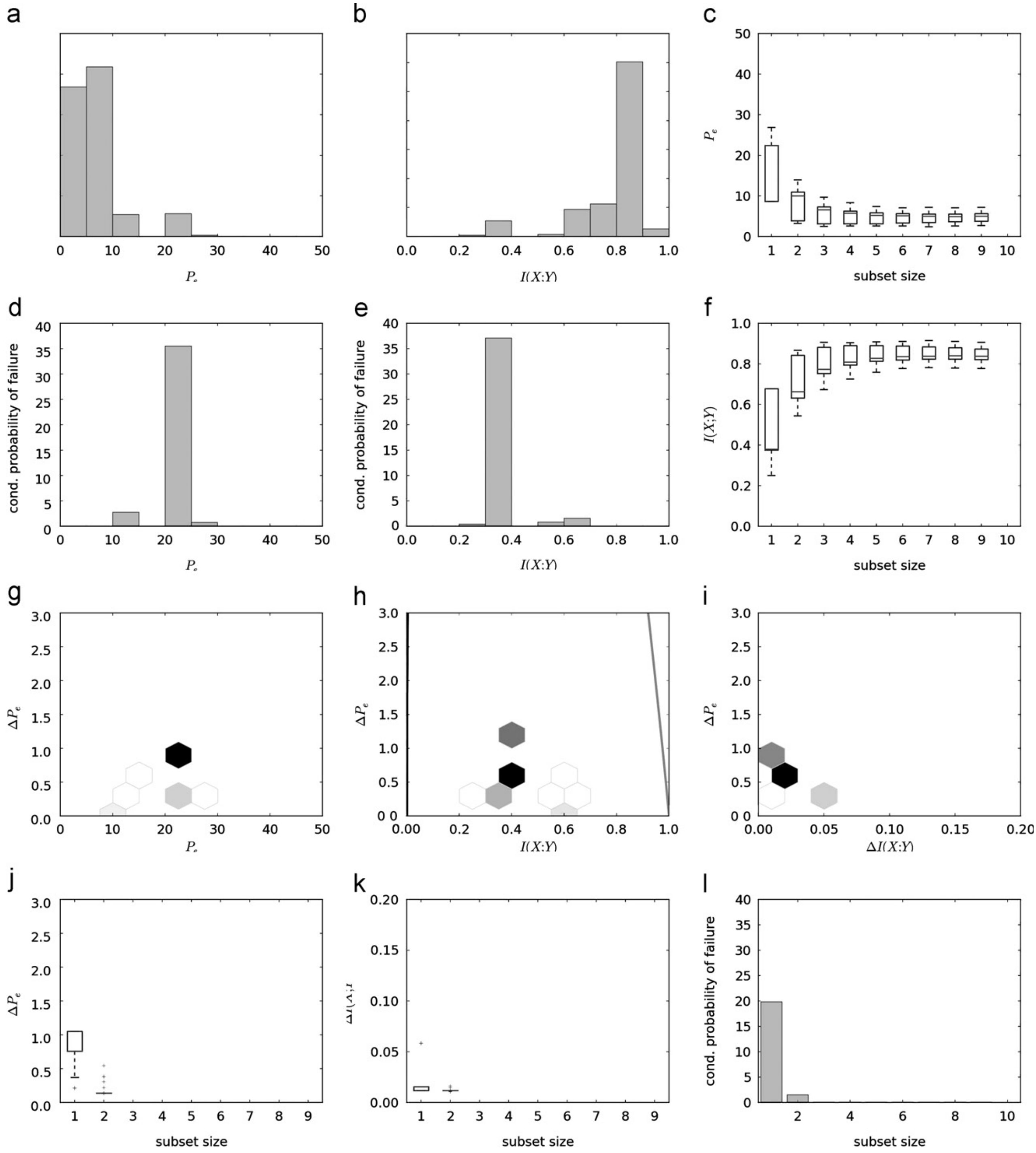

Fig. 14. Results of mutual information-based forward search with random subsets of 10 features for the Wallrobot dataset.

values and misclassification probabilities. Figs. 15(d) and (e) show a peak of probability of failure which corresponds to the dark hexagonal bin in Fig. 15(g)–(i), where 95% of the misclassification probability losses remain below 0.4%.

6.3. Discussion

The results of the above experiments on real-world datasets show that mutual information is more likely to fail in the first stages of the forward search. These situations correspond to intermediate values of mutual information. For the three datasets, misclassification probability losses remain in the order of the percent. This shows that mutual information failures do not have important consequences in practice. In each experiment, failures occur when the feature with the best mutual information and the feature with the best misclassification probability are close in terms of both mutual information and misclassification probability.

7. Meta-analysis of the experimental results

This section reviews and summarises the experimental results and the elements discussed in Sections 4–6, in order to extract several general conclusions. Firstly, the experiments show that

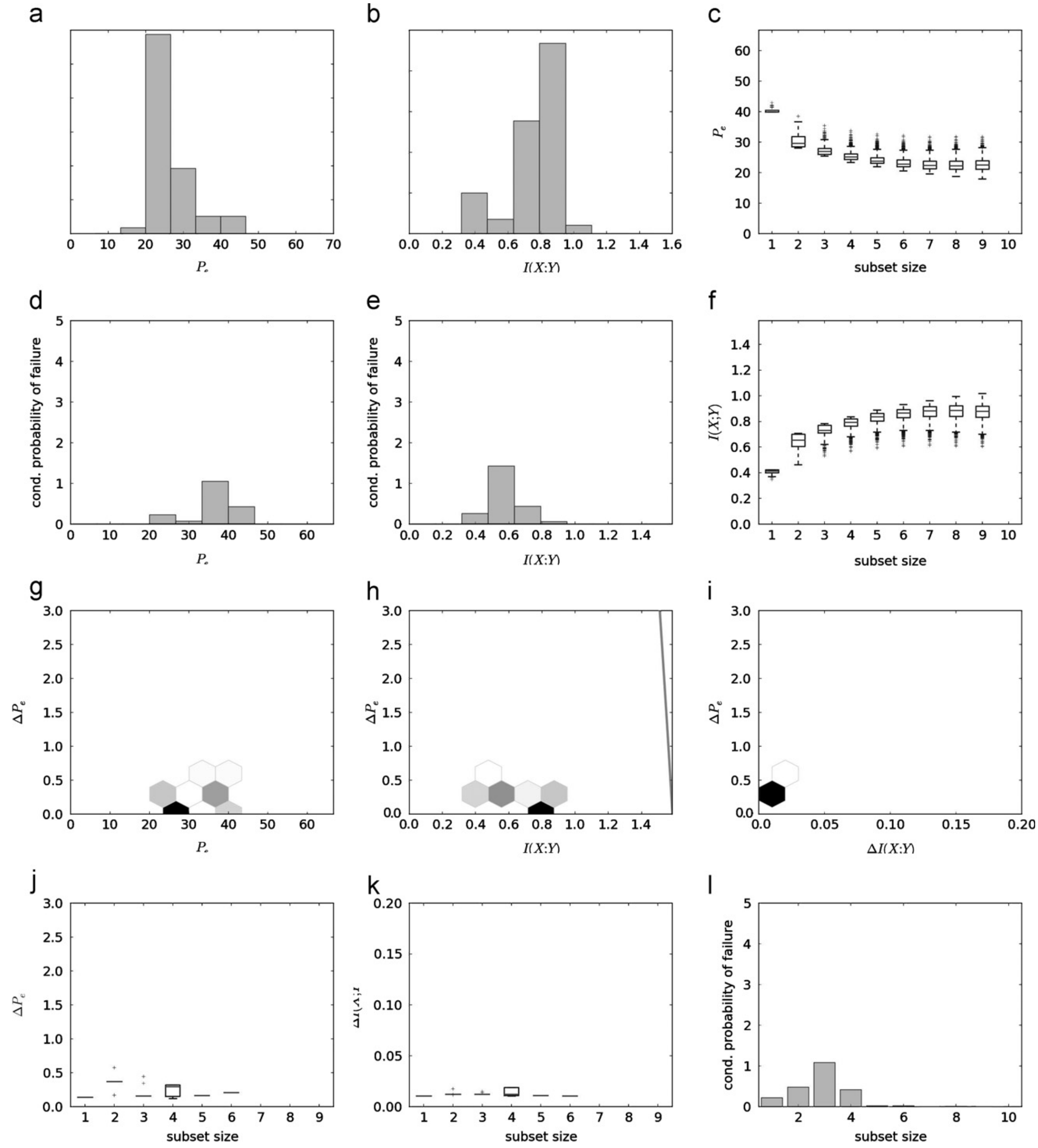

Fig. 15. Results of mutual information-based forward search with random subsets of 10 features for the Wave dataset.

mutual information can fail for feature selection in a wide range of artificial and real-world problems. However, the average percentage of failure is relatively small (often below 5%) and the misclassification probability loss remains in the order of a few percents. In particular, for the three real-world problems, the misclassification probability loss remains below 2% for 95% of the failures. Secondly, mutual information failures are more probable for intermediate values of mutual information. In forward selection, this case occurs in the first steps, when the feature subset size is still small. Hence, on a practical point of view, it could be a good idea to perform several backward steps after the first steps of the forward search, when the algorithm has reached a region where mutual information is more likely to be a reliable criterion for feature selection. Another effective option is to start a forward search with all combinations of 2 or 3 features (when computationally affordable) rather than with a single feature. Moreover, experiments suggest that the backward search algorithm could obtain more reliable feature subsets, since it directly starts in the region where mutual information is reliable and only reaches the dangerous zone after having found satisfying feature subsets. Thirdly, failures occur when comparing features which are close in terms of both mutual information and misclassification probability. Fourthly, in all experiments, the misclassification probability loss remains below the theoretical bound given in Section 3 and the Fano and Hellman–Raviv bounds are satisfied, what supports the validity of the experimental results.

8. Conclusion

This paper shows that in a classification context, mutual information is not always an optimal criterion to achieve feature selection, if the actual goal is eventually to minimize the probability of misclassification. Indeed, as it is first illustrated through a simple example, the Fano and Hellman–Raviv bounds do not guarantee such an optimality, contrary to what can be read in the literature. Extensive experiments on both continuous and discrete datasets confirm this fact and allow detecting the situations for which the mutual information criterion is the more likely to fail. It results that, taking some precautions and possibly adapting the search algorithm, mutual information remains a very interesting heuristic for feature selection.

Acknowledgments

The authors would like to thank the ESANN'12 reviewers and attendees for their fruitful discussions and comments on the subject of this paper, in particular Pierre Dupont, Amaury Lendasse, Fabrice Rossi and Jochen J. Steil.

References

[1] R.E. Bellman, Adaptive Control Processes—A Guided Tour, Princeton University Press, Princeton, New Jersey, USA, 1961.
[2] M. Verleysen, Learning high-dimensional data, Limitations Future Trends Neural Comput. 186 (2003) 141–162.
[3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[4] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (1994) 537–550.
[5] T.M. Cover, J.A. Thomas, Elements of Information Theory, 99th edition, Wiley-Interscience, 1991.
[6] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[7] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[8] F. Rossi, A. Lendasse, D. Francoi, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemom. Intell. Lab. Syst. 80 (2006) 215–226.
[9] B. Frénay, G. Doquire, M. Verleysen, On the potential inadequacy of mutual information for feature selection, in: Proceedings of ESANN 2012, 2012.
[10] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423 623–656.
[11] R. Fano, Transmission of Information: A Statistical Theory of Communications, The MIT Press, Cambridge, MA, 1961.
[12] J.W. Fisher, M. Siracusa, T. Kihn, Estimation of signal information content for classification, in: Proceedings of DSP/SPE 2009, 2009.
[13] M.E. Hellman, J. Raviv, Probability of error, equivocation and the Chernoff bound, IEEE Trans. Inf. Theory 16 (1970) 368–372.
[14] G. Brown, An information theoretic perspective on multiple classifier systems, in: Proceedings of MCS 2009, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 344–353.
[15] U. Ozertem, D. Erdogmus, R. Jenssen, Spectral feature projections that maximize Shannon mutual information with class labels, Pattern Recognition 39 (2006) 1241–1252.
[16] D. Francois, F. Rossi, V. Wertz, M. Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70 (7–9) (2007) 1276–1288.
[17] D.N.A. Asuncion, UCI machine learning repository, 2007 URL 〈http://www.ics.uci.edu/∼mlearn/MLRepository.html〉.
[18] L.F. Kozachenko, N. Leonenko, Sample estimate of the entropy of a random vector, Probl. Inf. Transm. 23 (1987) 95–101.

Benoît Frénay received the Engineer's degree from the Université catholique de Louvain (UCL), Belgium, in 2007. He is now a Ph.D. student at the UCL Machine Learning Group. His main research interests in machine learning include support vector machines, extreme learning, graphical models, classification, data clustering, probability density estimation and label noise.

Gauthier Doquire was born in 1987 in Belgium. He received the M.S. in Applied Mathematics from the Université catholique de Louvain (Belgium) in 2009. He is currently a Ph.D. student at the Machine Learning Group of the same university. His research interests include machine learning, feature selection and mutual information estimation.

Michel Verleysen received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an invited professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland) in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université ParisI-Panthéon-Sorbonne from 2002 to 2011, respectively. He is now a Full Professor at the Université catholique de Louvain, and Honorary Research Director of the Belgian F.N.R.S. (National Fund for Scientific Research). He is editor-in-chief of the *Neural Processing Letters* journal (published by Springer), chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning), past associate editor of the *IEEE Transactions on Neural Networks* journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He was the chairman of the IEEE Computational Intelligence Society Benelux chapter (2008–2010), and member of the executive board of the European Neural Networks Society (2005–2010). He is author or co-author of more than 250 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French, and of the "Nonlinear Dimensionality Reduction" book published by Springer in 2007. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.