Neurocomputing 90 (2012) 3–11

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Feature selection with missing data using mutual information estimators

Gauthier Doquire*, Michel Verleysen

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

ARTICLE INFO

Available online 22 March 2012

Keywords:
Feature selection
Missing data
Mutual information

ABSTRACT

Feature selection is an important preprocessing task for many machine learning and pattern recognition applications, including regression and classification. Missing data are encountered in many real-world problems and have to be considered in practice. This paper addresses the problem of feature selection in prediction problems where some occurrences of features are missing. To this end, the well-known mutual information criterion is used. More precisely, it is shown how a recently introduced nearest neighbors based mutual information estimator can be extended to handle missing data. This estimator has the advantage over traditional ones that it does not directly estimate any probability density function. Consequently, the mutual information may be reliably estimated even when the dimension of the space increases. Results on artificial as well as real-world datasets indicate that the method is able to select important features without the need for any imputation algorithm, under the assumption of missing completely at random data. Moreover, experiments show that selecting the features before imputing the data generally increases the precision of the prediction models, in particular when the proportion of missing data is high.

1. Introduction

Missing data are a very common problem which is important to consider in many data mining, machine learning or pattern recognition applications [1,2]. The reasons for which a value can be missing in a dataset are numerous. As an example, in industrial applications, data can be missing because of the dysfunction of an equipment or the insufficient resolution of a sensor device. In the socio-economic area, if data are collected from surveys, it is likely that some people will refuse to answer too personal questions about their income or their political opinions, leading to missing values in the datasets [3]. An other example is the medical domain where it is not always possible to conduct some experiments on certain patients; this can be due to their condition, the fact that they left the hospital or died or simply because of their age or their sex. Eventually, some data can simply be lost. These examples illustrate the different missingness mechanisms often considered in the literature [4].

Denote by M the indicator coding which values are observed and which are missing. In this paper, M is a matrix of the same size as the dataset and whose elements are 1 or 0, depending if the values are missing or not. Denote then the total dataset by $\Delta = \{\Delta_{obs}, \Delta_{mis}\}$ where Δ_{obs} is the observed part of the data and Δ_{mis} is the missing part.

The first popular assumption is that the data are missing at random (MAR), meaning that M (or its distribution) does not depend on the missing values

$$P(M|\Delta) = P(M|\Delta_{obs}). \tag{1}$$

This corresponds, in the medical example, to the fact that an experiment has not been carried out because the patient is (for instance) a woman, as indicated by a feature in the dataset.

As a special case of MAR data, if the missingness is furthermore independent of the observed values, the data are said to be *missing completely at random* (MCAR)

$$P(M|\Delta) = P(M). \tag{2}$$

This assumption is true for lost data or when a measure equipment unexpectedly stops working properly. It is also verified when one voluntarily does not ask all the participants of a test to undergo a huge number of experiments but randomly selects a few ones for each participant for time or cost reasons [5].

Eventually, when Eq. (1) does not hold, the data are *missing not at random* (MNAR). In this situation, the probability to observe a missing data thus depends on Δ_{mis} (and also possibly on Δ_{obs}); this can happen for example when people having a too low or too high income refuse to communicate it. MNAR missingness is a much harder problem to address than MAR or MCAR since the missingness distribution has to be modelled. In this paper the data will be assumed to be MCAR and experiments will be conducted accordingly.

* Corresponding author. Tel.: +32 10 47 81 33.
E-mail addresses: gauthier.doquire@uclouvain.be (G. Doquire), michel.verleysen@uclouvain.be (M. Verleysen).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.031

Even in the presence of missing values, the feature selection preprocessing step remains of great importance for many machine learning tasks, including regression and classification. As it is well known, feature selection aims at reducing the dimension of the original feature space by determining a (small) subset of the most relevant features for a given problem [6].

The benefits of feature selection can potentially be numerous. First, it can improve the performances of the prediction models by removing irrelevant and/or redundant features, and by preventing the models to suffer from the curse of dimensionality [7]. It can also help better understanding the problems by identifying the actually influent features (this is particularly important in the industrial and medical fields) and better interpreting the prediction models built. Due to these arguments, the areas for which feature selection has proven essential are numerous and include bioinformatics [8], text classification [9] or near infra-red spectra analysis [10] among many others.

Despite the fact that the presence of missing values in a dataset does not automatically lead to a great loss of information, it usually prevents one from using traditional data mining or machine learning tools. As an example, distances between observations, which are at the core of many algorithms, cannot be computed anymore. Specifically, most of the existing feature selection methods are designed to work with complete dataset and cannot be trivially adapted to handle problems with missing data. One obvious solution is of course to impute the data before applying classical procedures. However, such a procedure would add a bias whose effect on feature selection would be really hard to estimate; it is thus important in practice to select features independently of any imputation procedure.

The objective of this paper is precisely to propose a way to achieve feature selection when missing data are present, without the need for any imputation algorithm. Besides the motivation stated above, very few prediction models are able to handle missing data. Consequently, if one wants to achieve regression or classification with an incomplete dataset, it is very likely that an imputation phase will still be necessary after the feature selection step. In this case, it seems obvious that useless features could harm the imputation procedure. This will particularly be the case when distance-based imputation strategies (such as the very popular k nearest neighbors imputation and its variants) will be used.

To achieve imputation-free feature selection, the mutual information (MI) criterion will be employed. More precisely, it is proposed to estimate the MI directly from the dataset, adapting a recently introduced k-nearest neighbors based MI estimator [11]. As will be detailed later in this paper, this estimator offers very desirable properties for this task.

Preliminary ideas to perform feature selection in the presence of missing data were already published in [12]. This paper extends the methodology and sets up a sound experimental framework to assess the performances of the proposed algorithm.

The remaining of the paper is organized as follows. Section 2 draws a brief literature review about missing data analysis and feature selection. In particular, basic concepts about the MI are recalled and the MI estimator introduced in [11] is described. The proposed feature selection algorithm is introduced in Section 3. Section 4 is devoted to the presentation of experimental results while Section 5 concludes the work and gives some future work perspectives.

2. Missing data and feature selection: literature review

This section first gives a brief overview of the literature about the treatment of missing data in data mining as well as the feature selection problem. Basic definitions about MI are then recalled, before a nearest neighbors based MI estimator [11] is presented.

2.1. The missing data problem

As already stressed, the missing data problem is very common in machine learning, and has thus been widely studied in the literature, for a variety of purposes. Traditional approaches essentially include case deletion, missing values imputation and learning directly with missing data.

Case deletion, sometimes referred to as complete-case analysis, consists in considering in the analysis only the samples for which all the values are available, thus deleting the incomplete ones [13]. Obviously, this strategy can lead to a loss of information, especially when the proportion of missing data is high but can be used in combination with any algorithm designed for complete case problems.

On the other hand, imputation methods try to estimate the missing values, producing complete datasets. Traditional data mining or machine learning methods can then be applied. Lots of different imputation methods can be thought of and have been proposed in the literature, as will be discussed later in this work. Those techniques are extremely popular and have been used successfully, to name a few examples, in DNA microarray [14] and environmental [15] data analysis.

One of the most widely used imputation technique is the k-nearest neighbors imputation, which consist in finding, among the complete samples, the k-nearest neighbors of an incomplete data point. The missing values in this point are then replaced with an average of the values of its neighbors. The performances of this strategy are seriously limited when the proportion of missing data is high, and when only a few samples are complete. A simple improvement consists in looking also for incomplete neighbors of a sample, provided these neighbors are observed for the features missing in the sample. This strategy is called incomplete case k-nearest neighbors imputation (ICkNNI) [16].

Another simple imputation technique is to replace each missing value with the mean of the observed values for the corresponding feature. Other more sophisticated imputation methods have also been proposed, including an expectation maximization (EM) algorithm and regularized versions of it [17], as well as multiple imputation [18].

Eventually, methods have been developed that deal directly with missing data problems and do not require any imputation or deletion phase. As an example, in [19], the authors perform logistic regression with missing data by estimating conditional density functions using a Gaussian mixture model. For clustering purposes, [20] proposed the KSC algorithm, which encodes the partially observed features as a set of supplemental soft constraints.

2.2. The feature selection problem

Currently, three approaches are mainly followed to achieve feature selection: wrappers, filters and embedded methods.

Wrappers are directly based on the performances of a particular prediction model, and thus require the optimization of many of such models, which can be extremely time-consuming in practice. However, they are expected to lead to high prediction performances, precisely because they are designed to maximize the model performances [21].

On the contrary, filters search for the subset of features optimizing a criterion independently of any prediction algorithm. The most popular criteria used for feature selection are the correlation coefficient [22] and the mutual information (MI) [23,24] or other information-theoretic quantities [25]; many others have been proposed in the literature. Due to their simplicity and their rapidity,

filters are very popular in practice. It is the approach that will be followed in this work.

Eventually, embedded methods, which achieve simultaneously prediction and feature selection, have received a lot of attention recently, especially concerning extensions of the original LASSO [26,27].

Despite the importance of considering both the missing values and feature selection simultaneously, we have knowledge of only two works addressing this question. In [28], the authors try to infer the MI distribution using a second-order Dirichlet prior distribution to achieve feature selection. The developments are, however, limited to the classification problem. In [29], the authors propose to combine feature selection and imputation of missing values. However, the feature selection is only used to increase the performances of a k-nearest neighbors imputation algorithm and it is not clear if such a strategy can increase the performances of a prediction algorithm.

2.3. Mutual information: state of the art

Since its introduction by Shannon in 1948 [30], MI has been a very popular and successful criterion for feature selection (see, e.g., [31,32]). The reasons for this are essentially twofold. First, MI is able to detect non-linear relationships between variables, which is a strong advantage over, for example, the correlation coefficient which is limited to linear dependences. Then, MI can be naturally defined for groups of features; such a multivariate criterion can be very helpful in feature selection if many variables are redundant or if they are only jointly relevant to the output vector (as for the XOR problem).

2.3.1. Definitions

The MI is a symmetric measure of the dependence between two (groups) of random variables X and Y, considered to be continuous in this work. It can be defined in terms of another information-theoretic quantity, the entropy

$$H(X) = -\int p_X(\eta)\log p_X(\eta)\,d\eta, \tag{3}$$

where p_X is the probability density function (pdf) of X. The MI is then

$$I(X;Y) = H(Y) - H(Y|X), \tag{4}$$

with $H(Y|X)$ being the conditional entropy of Y given X.

Knowing that the entropy is a measure of the uncertainty about a random variable, Eq. (4) can be interpreted as follows. Let T be an output vector, whose values will have to be predicted for new samples based on an associated training set S (this is a simple supervised prediction problem). Let n be the number of samples ($T \in \Re^{n\times 1}$) and f be the original number of features in $S = (S^1 \dots S^f)$ ($S \in \Re^{n\times f}$). Then, if one tries to find a subset S_{sub} of the features in S maximizing $I(S_{sub};T)$, he actually tries to find a subset of features whose knowledge minimizes the uncertainty about T. This is obviously a quite natural criterion for feature selection.

An equivalent expression for the MI is

$$I(X;Y) = \iint p_{X,Y}(\eta,\zeta)\log\frac{p_{X,Y}(\eta,\zeta)}{p_X(\eta)p_Y(\zeta)}\,d\eta\,d\zeta. \tag{5}$$

This last equation can be seen as the Kullback–Leibler divergence between the joint distribution $p_{X,Y}$ and the product of the marginal distributions p_X and p_Y [33]. If the variables are independent $p_{X,Y} = p_X p_Y$ and $I(X;Y)=0$ as could already be deduced from Eq. (4).

Unfortunately, in practice, for real-world problems, none of the pdf p_X, p_Y or $p_{X,Y}$ in Eq. (5) are known, meaning that the MI cannot be directly computed but has to be estimated from the dataset.

2.3.2. Mutual information estimation

Mutual information estimation is a well-known and widely studied problem in the literature. Traditional approaches essentially consist in approximating the pdf by histograms or kernel based methods [34,24], before plugging the obtained quantities in Eq. (5) to obtain a MI estimation.

However, even if very popular, those methods are not likely to work well when working with high-dimensional groups of features. The reason is that the number of points needed to sample a space at a given resolution grows exponentially with the dimension of this space [7]. As the number of available data points is always limited in practice, if those samples are of high dimension, most of the boxes of an histogram will be empty, leading to very imprecise density estimations. Things will not be very different for kernel-based density estimators, since they are essentially smoothed histograms.

Nevertheless, considering a multivariate criterion for feature selection can be useful and a way to properly estimate the MI in high-dimensional spaces is thus needed. An estimator based on the notion of nearest neighbors has been introduced in [11]. This estimator has the crucial advantage that it does not directly require the estimation of any pdf, thus bypassing one of the most problematic issues in MI estimation. It has already been used successfully for feature selection in the no missing value case [10,35].

Let $\chi = \{x_i, i = 1\dots n\}$ and $\Upsilon = \{y_j, j = 1\dots n\}$ be the sets of realizations of the random variables X and Y, respectively. The estimator is based on the Kozachenko–Leonenko estimator of entropy [36]

$$\hat{H}(X) = -\psi(k) + \psi(n) + \log(c_d) + \frac{d}{n}\sum_{i=1}^{n}\log(\epsilon_X(i,k)), \tag{6}$$

where k is the number of nearest neighbors considered (a parameter of the estimator), d the dimensionality of the data, c_d the volume of a unitary ball of dimension d and $\epsilon_X(i,k)$ twice the distance from x_i to its kth nearest neighbor. Eventually, ψ is the digamma function defined as follows:

$$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)} = \frac{d}{dk}\ln\Gamma(k), \quad \Gamma(k) = \int_0^\infty u^{k-1}e^{-u}\,du. \tag{7}$$

Practically, the function ψ satisfies the following recursion: $\psi(u+1) = \psi(u) + 1/u$ and $\psi(1) = C$, $C = -0.5772\dots$ being the Euler–Mascheroni constant.

Based on (6), Kraskov et al. [11] derived two different estimators for regression problems. The most popular is

$$\hat{I}(X;Y) = \psi(n) + \psi(k) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}(\psi(\tau_x(i)) + \psi(\tau_y(i))), \tag{8}$$

where $\tau_x(i)$ is the number of points located no further than $\epsilon(i,k)$ from x_i and $\tau_y(i)$ is defined similarly in Υ. Here, $\epsilon(i,k)$ is defined as $\max(\epsilon_X(i,k),\epsilon_Y(i,k))$.

As can be seen, the estimator is only determined by the distance between the samples either in χ or in Υ. In practice the Euclidean distance is often used.

3. Feature selection with missing data

This section describes the proposed methodology to achieve feature selection in the presence of missing data. A feature selection algorithm is generally composed of a subset quality criterion, a search procedure to optimize it and a stopping criterion. In this work, we will essentially focus on the two first elements, leaving the stopping criterion for further considerations. In practice, the procedure can be stopped when a fixed number of features have been chosen. If a sufficient number of

samples is available, it is also possible to use a validation test to determine the optimal number of features. Another more sounded strategy, based on resampling methods, has also been proposed and used successfully for MI based feature selection [37].

3.1. Mutual information with missing data

For the reasons described above, the MI is the criterion chosen to achieve feature selection in this work. As already explained, the estimator (8) is only determined by the distances between each point and its k-nearest neighbors in both the X and the Y spaces.

Of course, the simplest strategy to compute these distances is to impute the missing values in order to obtain a complete dataset; the distances can then be directly obtained. As detailed in Section 2.1, imputation is extremely popular and many different strategies can be thought of.

However, to estimate the MI, the imputation of the missing values is not required since only the distance between samples is needed. Moreover, as detailed above, we would like the MI estimation to be independent of any imputation procedure. To this end, the partial distance strategy (PDS) will be used in this work. The quite simple idea behind PDS is to compute the distances based only on the available values, ignoring the missing ones in the analysis.

Consider two samples a, $b \in \Re^f$ and denote by M_a (M_b) the indices of the missing components for a (b). The distance between a and b as estimated by the PDS is

$$dist(a,b) = \sqrt{\frac{1}{f-|M_a \cup M_b|} \sum_{i \notin M_a \cap M_b} (a_i - b_i)^2}, \tag{9}$$

where $|\cdot|$ denotes the cardinality of a set. In this last equation, the estimated distance is thus normalized by the number of features (or dimensions) used to compute it. The objective is to make the range of ϵ_X and ϵ_Y comparable in (8). Indeed, in the feature selection context, Y is the random variable indicating the class; therefore its dimension is 1. Conversely, the dimension of X can be higher in practice. Without the proposed normalization, it is likely that ϵ_X would always be greater than ϵ_y (because it would be calculated using more dimensions) and the MI estimation would be harmed.

As a very simple example, the distance between the vectors [3 • 8 7 2] and [1 9 • 3 4], where • denotes a missing value, is computed by considering the first, fourth and fifth elements of each vector; it is equal to

$$\sqrt{\frac{(3-1)^2+(7-3)^2+(2-4)^2}{3}}.$$

Once the pairwise distances between samples have been obtained, the nearest neighbors can be determined in both the X and the Y spaces, and the MI can be estimated using (8).

PDS has already been used successfully in other domains such as clustering [38], inference of prediction problems [39] and self-organizing maps [40]. It is important to note, however, that the term distance has been used abusively in this section since the similarity measure defined by the PDS is not guaranteed to obey the triangle inequality.

3.2. Subset construction strategy

Once the subset quality criterion has been defined, a strategy to optimize it has to be chosen. In this paper, a greedy forward search procedure is employed, mainly because of its simplicity and its low computational cost. The idea of such a procedure is to start by selecting the feature whose MI with the output vector is the highest. Then, at each of the following steps, the feature whose addition to the already selected features produces the subset having the highest MI with the output vector is selected, hence the name greedy. The term forward comes from the fact that once a feature has been chosen, it is never removed from the set of selected features.

At each step of the feature selection algorithm, the dimension of the feature set whose MI with Y has to be estimated thus increases by one unity; the need for an estimator able to handle multidimensional vectors clearly appears.

Obviously, other choices of greedy search procedures could have been made as well. For example, a backward step, consisting in removing a variable if its removal increases the estimated MI, could have been added after each forward step described above [10]. Another idea would be to consider the backward elimination, whose principle is to start with all the features, and to remove them one at a time. Other more complex algorithms are also possible, including genetic algorithms or hill climbing strategies. See [41] for a more detailed overview of subset construction strategies.

3.3. Practical considerations

Before presenting experimental results in the following section, some practical details about the proposed methodology have to be discussed.

First, the MI estimator used in this work (8) has a parameter k corresponding to the number of nearest neighbors considered which needs to be determined. Following the recommendations in the original paper [11], a moderate value of 6 will be arbitrarily chosen in order to get a good tradeoff between variance and bias.

Then, it is obvious that the PDS distance strategy is not able to produce a distance estimate between two points if they have no common features for which their values are observed. In this case, the distance between the points is considered as infinite such that the points cannot be neighbors.

Eventually, before the MI estimation, the real-world datasets have to be normalized by removing the mean of each feature and dividing them by their standard deviation. This is because in the absence of any prior knowledge, we do not want that, due to their range, some features carry more weight than others in the computation of the pairwise distances between samples.

4. Results and discussions

This section illustrates the interest of the proposed feature selection methodology through various experimental studies, on both artificial and real-world datasets. The objective is to compare the performances of this new method with the strategy consisting in imputing the datasets before running the classical MI-based feature selection algorithm using the estimator (8) with the Euclidean distance. The section begins by the description of the datasets that will be used before the experimental setup is described. The results are then presented. Eventually, a discussion about the results concludes the section.

4.1. Datasets

Both artificial and real-world datasets are considered in this study.

4.1.1. Artificial datasets

Three artificial datasets are used to evaluate the performances of the proposed algorithms. The three datasets consist in 10 continuous features, uniformly distributed between 0 and 1. They are built such that only a subset of these 10 features are relevant to predict the associated continuous output.

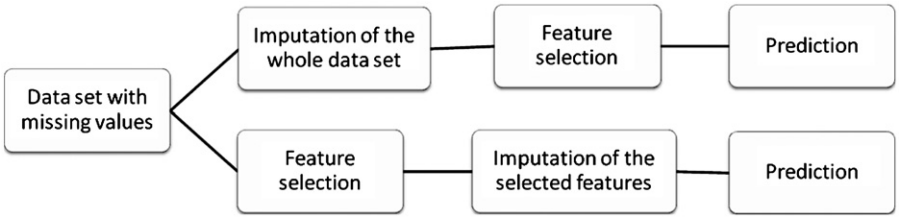

Fig. 1. Two different approaches to the regression problem with missing values.

The first artificial problem is derived from Friedman [42]; its output is defined as

$$Y_1 = 10\sin(X_1X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon, \quad (10)$$

where ϵ is a Gaussian noise with unit variance.

The second output is computed as

$$Y_2 = X_1X_2 + \sin(X_3) + 0.5X_4 + \epsilon, \quad (11)$$

while the third is built as

$$Y_3 = \cos(X_1X_2)\exp(X_3X_4) + \epsilon. \quad (12)$$

Obviously, only the five first features are useful to predict Y_1 while only the four first ones are needed to predict Y_2 and Y_3. The quality of a feature selection method can thus be studied through its ability to detect those relevant features.

4.1.2. Real-world datasets

Four real-world datasets are considered; two are high-dimensional while the other two have a more reasonable number of features around 15. The first one is the Delve census dataset, available from the University of Toronto[1] for which only the 2048 first entries are kept. The first 1500 entries are used as training set and the remaining ones as the test set. The objective is to predict, based on demographic features, the median price of houses in a region. The dimension of the dataset is 104.

The second dataset is the Nitrogen dataset, containing only 141 spectra discretized at 1050 different wavelengths. The goal is to predict the nitrogen content of a grass sample. The data can be obtained from the Analytical Spectroscopy Research Group of the University of Kentucky.[2] As a preprocessing, each spectrum is represented using its coordinates in a B-splines basis, in order to reduce the amount of features to 105; see [43] for details. The training set is composed of 105 data points, while the remaining 36 constitute the test set.

The third dataset is the Housing dataset, available from the UCI Machine Learning repository.[3] The goal is to predict the values of houses using 13 attributes. From the 506 instances, 337 are kept for training the model while the 169 remaining ones constitute the test set.

Finally, experiments are also carried out on the Mortgage dataset which contains 16 features. From the 1049 data points available, the 700 first ones are used to train the models. The dataset can be downloaded directly from the website of the Federal Reserve Bank of Saint-Louis.[4]

4.2. Experimental setup

As detailed above, the objective of this section is to compare the proposed approach with the approach consisting in first imputing the data before doing feature selection. To this end, two imputation strategies will be used: the ICkNNI with 10 neighbors and a regularized version of the EM algorithm [17]. It is important to note that the ICkNNI imputation method fails to produce an estimation of the missing values if, for a given sample whose value is missing for some features, no neighbors for which those features are observed can be found; the mean value of the feature is used instead in such situations.

4.2.1. Artificial problems

For artificial problems, the relevant features are known in advance. No prediction model is thus needed to asses the quality of the approaches. Three strategies are then compared: the proposed approach, feature selection after imputation by the EM algorithm and feature selection after imputation by ICkNNI. For each problem (10)–(12), 50 datasets of sample size 100 have been generated and the outputs have been built accordingly. Each dataset has been filled randomly with 10% and 40% of missing values, respectively. The different MI estimators have then been used to rank the features containing missing values.

4.2.2. Real-world problems

For real-world datasets, since the relevant features are not known in advance, the criterion of comparison between feature selection techniques will be the root mean square error (RMSE) of a prediction model using the selected features. In practice, very little work has been done to design prediction models able to handle missing values, with a few exceptions for classification problems (see e.g. [44]). Because of this, an imputation phase is still necessary after the feature selection step if one intends to build a prediction model after having used the PDS based strategy to select relevant features. Conversely, if the data have already been imputed, any classical prediction model can be built based on the selected features. Fig. 1 summarizes the two different approaches to the prediction problem with missing data compared in this paper. Two different regression models are used: the well known k-nearest neighbors predictor, with $k=10$, and a Radial Basis Functions Network model (RBFN). The parameters of the RBFN have been optimized according to [45].

In each of the four complete datasets, 1, 5, 10 and 20% of missing values are introduced. This generation is repeated 10 times, producing 40 new datasets. The feature selection procedure is then conducted, either directly with the estimator described in Section 3.1 or after imputation of the data, using the classical estimator (8). When the PDS-based MI estimation is used, the selected features are imputed; a prediction model is then built (see the lower part of Fig. 1).

4.2.3. Feature subset construction strategies

In addition to the forward search described in Section 3.2, feature ranking is also considered in this section. Indeed, sometimes, one is only interested in ranking the features according to their individual MI with the output Y. Even if this strategy does not take into account any possible redundancy between features or the fact that some features can only be jointly relevant, it has the advantage of being faster than the forward strategy and can be sufficient in some

[1] http://www.idrc-chambersburg.org/index.html.
[2] http://kerouac.pharm.uky.edu/asrg/cnirs/.
[3] http://archive.ics.uci.edu/ml/index.html.
[4] http://www.stls.frb.org/fred/data/zip.html.

situations. Moreover, a multivariate MI estimation, even using the nearest neighbors based estimator, will require more data points than a one-dimensional one, those data points being not always available. It is important to note that in the context of ranking procedures, the PDS approach is equivalent to the classical estimation of the MI based only on the samples for which the value of the feature is not missing. The same consideration is also true for the first step of the forward search procedure.

The ranking procedure will only be considered for the artificial problems, while the forward search will be conducted on both the artificial and the real-world problems. Indeed, a ranking procedure is likely to fail on the real-world problems, as many features are redundant, especially for the Nitrogen dataset. Moreover, the main objective of this paper is to assess the possibility of performing multivariate feature selection.

4.3. Results

The results obtained for the three experimental frameworks described above are now presented.

4.3.1. Ranking procedure for the artificial problems

Fig. 2 shows the average MI for each individual feature obtained over the 50 repetitions of the experiment. In order to assess the added value of the proposed methodology in each experiment individually, Table 1 shows the percentage of experiments for which five, four, three, two and one relevant features have, respectively, been ranked in the five first positions for the Y_1

Table 1
Percentage of experiments for which a certain number of relevant features (second line of the table) are ranked in the five first positions for the Y_1 problem with 10 and 40% of missing data.

Imputation strategy	10% of missing data					40% of missing data				
	5	4	3	2	1	5	4	3	2	1
PDS	14	49	30	6	1	8	40	43	9	0
Regularized EM	0	1	6	30	49	0	1	23	45	28
ICkNNI	0	1	5	39	43	0	0	9	43	40

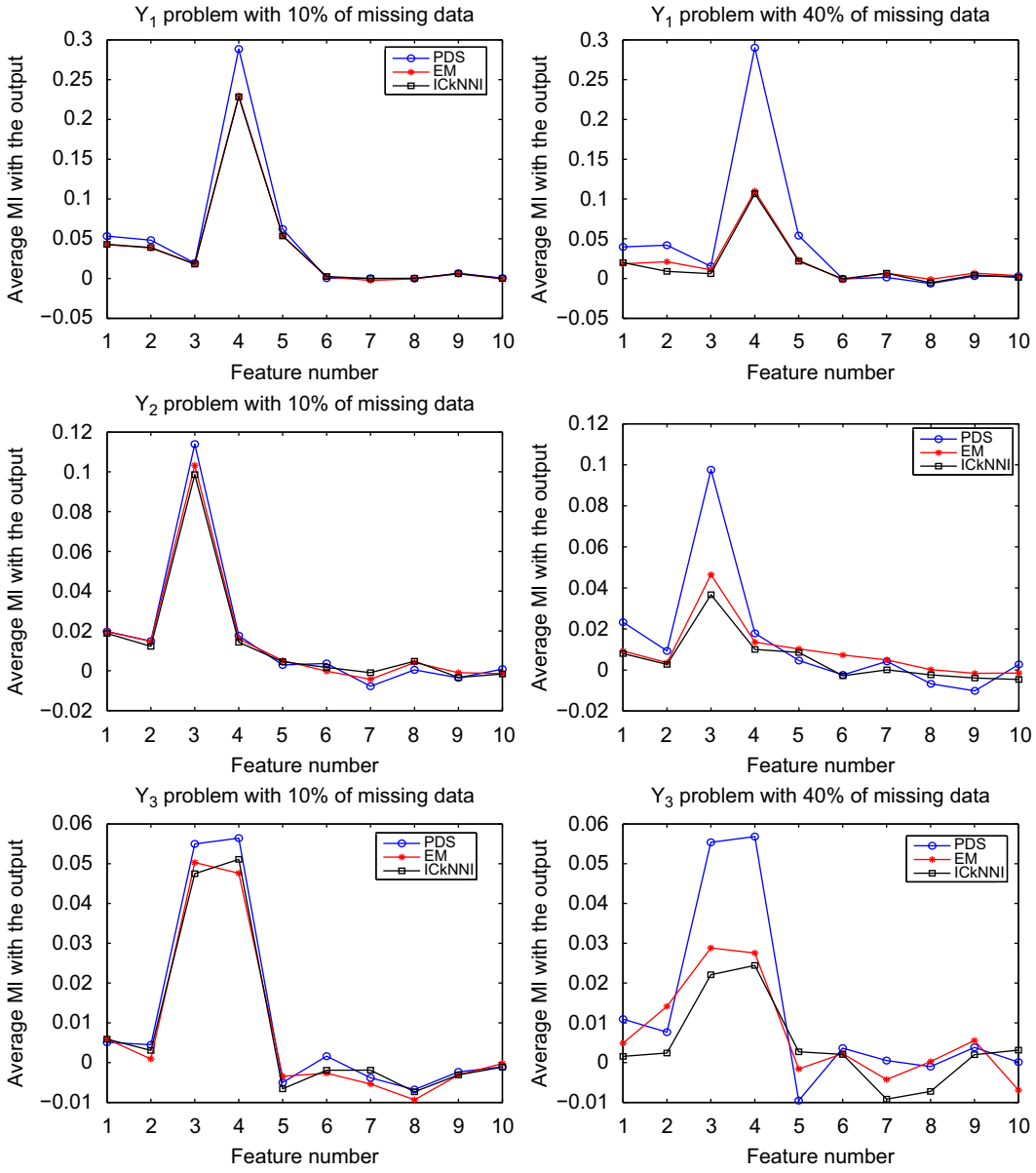

Fig. 2. Average MI between each individual feature and, from top, to bottom Y_1, Y_2, Y_3 with 10% (left) and 40% (right) of missing values; ○, PDS; □, ICkNNI; ★, regularized EM imputation.

problem. Results, which are similar for the two other artificial problems, are not shown for space reasons.

4.3.2. Forward procedure for the artificial problems

A forward procedure has been run on the datasets and stopped when the actual number of relevant features had been selected, i.e. five for the first problem, four for the last two ones. The percentages of relevant features among all the features selected by each of the strategies have then been computed. Table 2 summarizes the average performances of the three compared approaches when the output to predict is Y_1. For the two other problems, all methods always select the relevant features first. To complete those results, we show in Table 3 the percentage of experiments for which all relevant features have been selected.

4.3.3. Forward procedure for the real-world problems

For real-world problems, the number of selected features is arbitrarily limited to half the number of original features, with a maximum of 25. This allows us to see which methods quickly detect the most relevant features. Moreover, in [37], the authors propose an efficient stopping criterion using resampling methods for MI based forward feature selection. In [37], three out of the four datasets used in the present paper are also considered and the number of features selected is 8 or 9, 25 and 4 for the Delve census, the Nitrogen and the Housing dataset, respectively. Even if the criterion in [37] is designed for complete datasets, it indicates that only a small number of features is generally needed to obtain good prediction performances.

Table 4 shows the average best prediction performances obtained over the 10 incomplete datasets, together with the standard deviation, for the four missing data rates. EM or ICkNNI means that the data have first been imputed using the regularized EM algorithm or the ICkNNI. PDS then EM or ICkNNI means that the feature selection has been achieved without any imputation and indicates which imputation method has been used before building the prediction model. Cases for which one of the two strategies (imputation before or after feature selection) performs significantly better than the other one according to a paired test with a 95% confidence level are shown in bold.

Table 2
Average percentage of relevant features selected on the Y_1 artificial problem.

% of missingness	PDS	Regularized EM	ICkNNI
10	100	100	100
20	96,80	94,40	93,60
30	94,80	94	94

Table 3
Percentage of correct selection of the relevant set of features by a forward search procedure for the Y_1 artificial problem.

% of missingness	PDS	Regularized EM	ICkNNI
10	100	100	100
20	84	70	62
30	66	46	44

4.4. Analysis of the results

The performances of the compared approaches presented in Section 4.3 are now discussed in details.

4.4.1. Ranking procedure for the artificial problems

Results clearly show, in this configuration, the advantage of the PDS. More precisely, the results are quite similar when only 10% of missing values are missing (the left column of Fig. 2); in this situation, the three approaches are all able to discriminate between the irrelevant and the relevant features, giving the latest a higher value of MI. However, when the proportion of missing values is raised to 40%, it clearly appears that imputation based methods are not able anymore to clearly decide whether or not a feature is relevant to the output. Indeed, the difference between the values of the MI for the relevant and the irrelevant features is greatly reduced. As an example, for the second problem, it appears that the MI between feature 5 and the output is greater than the MI between the output and feature 1 or 2 when both imputation techniques are used. As another example, when the third problem is considered, imputation-based feature selection techniques indicate that feature 9 is more relevant to Y_3 than feature 1. Thus, such strategies could easily lead a user to wrong conclusions about the importance of each feature in a regression

Table 4
Best RMSE reached by a RBFN and a KNN prediction models on four real-world datasets for various rates of data missingness. Significantly better results for one of the two compared approaches are shown in bold.

Dataset	%	RBFN				KNN			
		EM	PDS then EM	ICkNNI	PDS then ICkNNI	EM	PDS then EM	ICkNNI	PDS then ICkNNI
Delve	1	1,9374 ± 0,0904	1,8922 ± 0,0543	**1,5976 ± 0,1112**	1,9023 ± 0,1106	2,1075 ± 0,4644	2,0790 ± 0,2827	2,0846 ± 0,2778	1,9263 ± 0,6923
	5	2,1483 ± 0,0526	**1,9114 ± 0,0730**	2,7002 ± 0,2820	**2,2726 ± 0,0985**	2,7126 ± 0,5282	2,4226 ± 0,5564	2,8999 ± 0,4434	**2,0288 ± 0,7085**
	10	2,4166 ± 0,3927	**1,9725 ± 0,1132**	3,4510 ± 0,1469	**2,6745 ± 0,2521**	2,4421 ± 0,3269	2,6606 ± 0,2374	3,1966 ± 0,1442	**2,3994 ± 0,2787**
	20	2,6002 ± 0,1262	**2,1744 ± 0,1262**	4,5044 ± 0,2578	**2,8726 ± 0,1439**	3,5703 ± 0,5510	3,6363 ± 0,4690	4,5026 ± 0,5780	**3,4126 ± 0,3856**
Nitrogen	1	0,6871 ± 0,0296	0,6796 ± 0,0297	0,7068 ± 0,0838	0,7052 ± 0,0419	0,7573 ± 0,0072	0,7541 ± 0,0323	0,7384 ± 0,0439	0,7330 ± 0,0303
	5	0,7297 ± 0,0079	0,7247 ± 0,0079	0,7921 ± 0,0213	0,7210 ± 0,2420	0,7566 ± 0,0098	0,7383 ± 0,0268	0,7362 ± 0,0455	0,7282 ± 0,0253
	10	0.7529 ± 0.0537	0.7254 ± 0.0232	1.0198 ± 0.1252	**0.7582 ± 0,0293**	0,7707 ± 0,0612	0,7502 ± 0,0277	0,8669 ± 0,0710	**0,7546 ± 0,0322**
	20	0,7728 ± 0,0417	0,7683 ± 0,0375	0,8309 ± 0,0924	0,8213 ± 0,3390	0,7722 ± 0,0498	0,7618 ± 0,0261	0,7741 ± 0,0274	0,7728 ± 0,0356
Housing	1	7,7308 ± 0,6103	7,6343 ± 0,7922	7,7187 ± 0,4537	7,6061 ± 1,2242	2,9366 ± 0,1277	2,8252 ± 0,0690	2,8896 ± 0,1783	2,8522 ± 0,0590
	5	8,1533 ± 0,9341	**7,0395 ± 0,2474**	7,0163 ± 0,7397	6,9950 ± 0,4565	3,1365 ± 0,1341	3,1083 ± 0,1548	3,9726 ± 0,2323	**2,9732 ± 0,1551**
	10	7,6845 ± 0,3980	7,3737 ± 0,2737	7,8922 ± 2,1218	7,4886 ± 0,5520	3,5232 ± 0,3061	3,4998 ± 0,3617	4,4350 ± 0,2238	**3,4646 ± 0,1942**
	20	8,3647 ± 0,4642	**7,6461 ± 0,4159**	9,3573 ± 0,9904	**7,6116 ± 0,3985**	3,6588 ± 0,5177	3,7271 ± 0,3508	6,0218 ± 1,0507	**4,0720 ± 0,3173**
Mortgage	1	0,1819 ± 0,0692	0,1517 ± 0,0521	0,1714 ± 0,0919	0,2184 ± 0,098	0,3709 ± 0,0193	0,3700 ± 0,0149	0,3515 ± 0,0401	0,3571 ± 0,0250
	5	0,2680 ± 0,0750	0,2668 ± 0,0646	0,3568 ± 0,1576	0,2664 ± 0,1344	0,3986 ± 0,0640	0,3545 ± 0,0394	0,4339 ± 0,0227	**0,3778 ± 0,0595**
	10	0,3068 ± 0,0757	**0,1901 ± 0,0624**	1,006 ± 0,463	**0,2556 ± 0,1106**	0,3908 ± 0,0337	0,3519 ± 0,0388	0,4950 ± 0,0986	**0,3589 ± 0,0428**
	20	0,3964 ± 0,1784	**0,2244 ± 0,0722**	1,6213 ± 0,284	**0,5071 ± 0,2719**	0,4854 ± 0,0501	**0,3680 ± 0,0686**	0,8355 ± 0,0946	**0,4057 ± 0,0435**

problem. On the contrary, when the PDS is used, the average MI is always higher for the relevant features than for the irrelevant ones.

Table 1 confirms these claims as the PDS-based strategy obviously leads to the selection of relevant features in a much bigger proportion of the experiments. As an example, for both 10% and 40% of missing data, at least three relevant features are chosen in more than 90% of the experiments.

4.4.2. Forward procedure for the artificial problems

The differences in performance, shown in Table 2, are not huge for the forward search procedure on the artificial datasets. However, the simple PDS-based strategy leads to the best results, confirming what had been observed above for the ranking procedure. Moreover, the results obtained are very encouraging showing that the method was effectively able to detect the most useful features in a very large majority of the experiments carried out. To support these claims, Table 3 shows that the percentage of experiments for which all relevant features have been selected is obviously higher for the proposed approach when the proportion of missing data grows; the interest of the PDS-based algorithm is clearly confirmed when both 20% and 30% of the data are missing.

4.4.3. Forward procedure for the real-world problems

The interest of the proposed approach on the real-world problems is clearly demonstrated by the results. Indeed, it leads to better average regression performances than the imputation-based feature selection techniques in all but five of the 64 cases. Moreover, it seems much less sensitive to the increase of the percentage of missing values. As an example, for the Mortgage dataset, Table 4 shows a reduction of the mean RMSE by more than 25% for ICkNNI imputation and by more than 50% for the regularized EM imputation with both prediction models when 20% of the values are missing. More generally, the RBFN model always performs better with the PDS-based method when 5% of values at least are missing. When 1% of data are unavailable, it is only less efficient for the ICkNNI imputation on the Mortgage and on the Delve datasets.

The results obtained with the KNN predictor confirm the good behavior of the proposed approach: results are always better with PDS when using ICkNNI imputation, and better in 13 out of 16 cases when using the regularized EM imputation.

As can further be seen in Table 4, the results obtained with the proposed strategy are statistically significantly better in 26 out of the 64 cases; on the contrary, imputing before doing feature selection only leads once to a significantly better prediction performance. The observations are in good agreement with the previous considerations and clearly establish the advantage of the proposed methodology.

4.5. Discussion

Even if the obtained results are clearly in favor of the PDS-based approach, some points still need to be discussed.

As mentioned above, parameter k of the Kraskov estimator has arbitrarily been set to 6. In practice this parameter plays a role in the bias-variance trade-off of the estimation. Experiments with different values of k chosen in a reasonable range have produced very similar results in our simulations. However, in some cases, results could vary and the value of k could then be important to set correctly. One way to proceed is to use resampling methods and the permutation test [37], which have proven to be successful for problems without missing data.

Then, the proposed methodology potentially suffers from a few drawbacks. First, as it is the case of most methods dealing with missing data, the performances of the algorithm are likely to decrease if the rate of missing data becomes very high. More precisely, the methodology is theoretically not limited to a certain percentage of missing values. However, for high rates of missingness, the first steps of the forward procedure could be achieved based only on a very few number of data points, leading to uncertain results. One possible solution would be to rather use a backward search strategy, or to begin the forward procedure with groups of two or three features instead of singletons; in this way, the whole procedure would not be affected by a bad choice of the first few features.

Closely related, the second problem is that there is currently no possible way to determine whether the estimated MI values are meaningful. In other words, the proposed methodology can potentially estimate the MI for any rate of missing value; however, once a given rate of missingness will be reached, the estimated values will probably not reflect the true dependencies anymore because the estimator will be given too few information. Knowing whether or not a value is relevant is obviously of great importance for feature selection. Here again, one solution could be found in the permutation test. Comparing the obtained estimations of the MI with the values of the MI between a random vector and the output could give valuable information on whether or not the estimations correspond to true dependencies.

5. Conclusions and future work

This paper proposes an approach to the feature selection problem for datasets where values are missing. The method is based on the concept of mutual information which has been widely used to achieve feature selection for complete datasets. To this end, a recently introduced MI estimator is adapted to handle missing data using the partial distance strategy and a greedy forward search is used to construct an optimal feature subset. One of the main advantages of the proposed feature selection algorithm is that it does not require any prior imputation of the data. On the contrary, traditional approaches consist in imputing the missing values before using any existing feature selection algorithm and are thus strongly dependent of the chosen imputation strategy.

Experimental results have first showed on three artificial datasets that the proposed algorithm is effectively able to identify the features relevant to a problem and that imputing the missing values leads in general to worse performances. Then, when working with real-world datasets, experiments indicate that imputing the missing values after the feature selection step generally allows us to build more precise prediction models, especially when the proportion of missing data is high.

As all the developments presented in this work could as well be applied to the classification-specific MI estimator introduced in [46], future work could be focused on experiments for this particular class of problems. Eventually, since the similarity measure defined by the PDS is not a metric, the effects of such a measure on the MI estimator should also be studied.

Acknowledgments

Gauthier Doquire is funded by a Belgian F.R.I.A. grant.

References

[1] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psycholog. Methods 7 (2) (2002) 147–177.
[2] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd ed., Wiley-Interscience, 2002.

[3] G. King, J. Honaker, A. Joseph, K. Scheve, Analyzing incomplete political science data: an alternative algorithm for multiple imputation, Am. Polit. Sci. Rev. 95 (2000) 49–69.
[4] D.B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592. http://dx.doi.org/10.1093/biomet/63.3.581.
[5] A.C. Acock, Working with missing values, J. Marriage Family 67 (2005) 1012–1028.
[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[7] M. Verleysen, Learning high-dimensional data, Limit. Future Trends Neural Comput. 186 (2003) 141–162.
[8] Y. Saeys, I.n. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.
[9] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 412–420.
[10] F. Rossi, A. Lendasse, D. Francois, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemomet. Intell. Lab. Syst. 80 (2006) 215–226.
[11] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[12] G. Doquire, M. Verleysen, Mutual information for feature selection with missing data, in: Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2011), 2011.
[13] J.R. Quinlan, Unknown attribute values in induction, in: Proceedings of the Sixth International Machine Learning Workshop, Morgan Kaufmann, 1989, pp. 164–168.
[14] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics (Oxford, England) 17 (6) (2001) 520–525.
[15] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmos. Environ. 38 (18) (2004) 2895–2907.
[16] J.V. Hulse, T.M. Khoshgoftaar, Incomplete-case nearest neighbor imputation in software measurement data, in: Proceedings of the IEEE International Conference on Information Reuse and Integration, IRI 2007, 2007, pp. 630–637.
[17] T. Schneider, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, J. Clim. 14 (5) (2001) 853–871.
[18] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York, 1987.
[19] D. Williams, X. Liao, Y. Xue, L. Carin, Incomplete-data classification using logistic regression, in: In ICML, ACM Press, 2005, pp. 972–979.
[20] K. Wagstaff, Clustering with missing values: no imputation required, in: Proceedings of the Meeting of the International Federation of Classification Societies, 2004, pp. 649–658.
[21] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[22] M. Hall, Correlation-based Feature Selection for Machine Learning, Ph.D. Thesis, University of Waikato, 1999.
[23] 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 (8) (2005) 1226–1238.
[24] R. Steuer, J. Kurths, C.O. Daub, J. Weise, J. Selbig, The mutual information: detecting and evaluating dependencies between variables, Bioinformatics 18 (Suppl. 2) (2002) S231–S240.
[25] G. Brown, A new perspective for information theoretic feature selection, J. Mach. Learn. Res.: Proc. Track 5 (2009) 49–56.
[26] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. B 67 (2005) 301–320.
[27] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R Stat. Soc., Ser. B 68 (2006) 49–67.
[28] M. Zaffalon, M. Hutter, Robust feature selection by mutual information distributions, in: Proceedings of the 18th International Conference on Uncertainty in Artificial Intelligence (UAI-2002), Morgan Kaufmann, 2002, pp. 577–584.
[29] P. Meesad, K. Hengpraprohm, Combination of knn-based feature selection and knn-based missing-value imputation of microarray data, in: Proceedings of the 2008 Third International Conference on Innovative Computing Information and Control, IEEE Computer Society, Washington, DC, USA, 2008, pp. 341–344.
[30] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[31] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (1994) 537–550. http://dx.doi.org/10.1109/72.298224.
[32] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[33] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[34] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076.
[35] F. Rossi, D. François, V. Wertz, M. Verleysen, Fast selection of spectral variables with b-spline compression, Chemomet. Intell. Lab. Syst. 86 (2) (2007) 208–218.
[36] L.F. Kozachenko, N. Leonenko, Sample estimate of the entropy of a random vector, Prob. Inf. Transm. 23 (1987) 95–101.
[37] 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.
[38] R.J. Hathaway, J.C. Bezdek, Fuzzy c-means clustering of incomplete data, IEEE Trans. Syst. Man Cybern. B: Cybern. 31 (5) (2001) 735–744.
[39] R.J. Almeida, U. Kaymak, J.M.C. Sousa, A new approach to dealing with missing values in data-driven fuzzy modeling, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ), 2010, pp. 1–7.
[40] A. Sorjamaa, A. Lendasse, E. Séverin, Combination of SOMs for fast missing value imputation, in: Proceedings of MASHS, Models and Learnings in Human and social Sciences, 2010.
[41] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer-Verlag, Inc., New York, 2006.
[42] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1) (1991) 1–67.
[43] F. Rossi, N. Delannay, B. Conan-Guez, M. Verleysen, Representation of functional data in neural networks, Neurocomputing 64 (2005) 183–210. (Trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks 2004).
[44] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor, Handling missing values in support vector machine classifiers, Neural Networks 18 (5–6) (2005) 684–692.
[45] N. Benoudjit, M. Verleysen, On the kernel widths in radial-basis function networks, Neural Process. Lett. 18 (2) (2003) 139–154. doi:http://dx.doi.org/10.1023/A:1026289910256.
[46] V. Gomez-Verdejo, M. Verleysen, J. Fleury, Information-theoretic feature selection for functional data classification, Neurocomputing 72 (16–18) (2009) 3580–3589.

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

Michel Verleysen was born in 1965 in Belgium. He received the MS and the PhD degrees in Electrical Engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (E.P.F.L.), Switzerland in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1-Panthéon-Sorbonne in 2002–2004. He is now a Research Director of the Belgian Fonds National de la Recherche Scientifique (F.N.R.S.) and Lecturer at the Université catholique de Louvain. He is Editor-in-Chief of the Neural Processing Letters journal, Chairman of the Annual European Symposium on Artificial Neural Networks (ESANN) Conference, 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 is the author or the co-author of about 200 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. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, non-linear statistics, adaptive signal processing, and high-dimensional data analysis.