Weaning Failure Prediction from Heterogeneous Time Series using Normalized Compression Distance and Multidimensional Scaling

J.M. Lillo-Castellano, I. Mora-Jiménez, J.L. Rojo-Álvarez and A. Algora-Weber

***Abstract*—Scientific evidence has shown that a failed weaning, defined as the process of gradual reduction in the level of mechanical ventilation support, increases the risk of death in prolonged mechanical ventilation patients. Different methods for weaning outcome prediction have been proposed, using variables and time series from the monitoring systems, however, monitored data are often non-regularly sampled, hence limiting its use in conventional automatic detection systems. In this work, we propose the joint use of two statistical data techniques, the Normalized Compresion Distance (NCD) and the Multidimensional Scaling (MDS), to deal with the data heterogeneity in the monitoring systems for failure weaning prediction. NCD technique determines a similarity measure between two sequences from compression length and entropy principles, whereas MDS provides each sequence with a point in an N-dimensional space suitable for training a classifier for weaning prediction. A total of 104 weaning events were collected in 253 patients under mechanical ventilation from Intensive Care Unit of Hospital Universitario Fundación Alcorcón; for each weaning event, 20 time series (TS), 15 clinical laboratory parameters (CLP) and 12 general descriptors (GD) were collected during 48 hours previous to the weaning event. Only 18 TS could be considered as candidates to the classifier input space, and one of them, diastolic blood preassure, reached significant results by itself, yielding 90.4% of accuracy prediction. These results show that weaning prediction systems can be designed with NCD and MDS, providing a compact input space to the classifiers.**

***Index Terms*—Weaning, Normalized Compression Distance, Multidimensional Scaling, Partial Least Squares.**

I. Introduction

IN daily's routine of the Intensive Care Unit of a hospital (ICU), patients are often under mechanical ventilation, this is, a mechanically assisted method that replaces or collaborates with the spontaneous breathing of a patient with respiratory problems. The process of discontinuing mechanical ventilation is usually called *weaning*, and it consist of a gradual removal in the level of respiratory support under mechanical ventilation [1]. Although current mechanical ventilators are sophisticated devices which are capable of stabilizing the respiratory conditions of a patient, the decision about the exact time of withdrawal mechanical support (*extubation*) is under the responsibility of a physician and has several problems. On the one hand, a premature extubation can increase the complication rate in the patient, causing difficulty in reestablishing artificial airways and compromising gas exchange [2]. On the other hand, an unnecessary delay in the discontinuation of mechanical ventilation brings other problems, such as pneumonia or airway trauma, as well as an increase in the economic cost to the hospital [3]. Hence, two main questions have to be taken into account in the weaning setting, specifically, how can the physician better decide the best moment for a patient to be extubated, and which information can be used to support this decision.

Nowadays, physicians using their knowledge and own experience can identify when the patient is ready for beginning his weaning from mechanical ventilation and which method is the most appropiate for it, as several procedures are available [4]. Actually, the most used methods for weaning consist on assesing the patient's respiratory status by observing either his spontaneous breathing through a T-Tube circuit (T-Tube Test) or his breathing while is assisted by a low pressure support. If the patient tolerates the test (true-positive result) and if the physician considers the weaning appropriate, then the patient is extubated. An example of an alternative method is shown in [5], where a therapist-implemented protocol was used to extubate patients from prolonged mechanical ventilation for reducing the time to weaning, though the disconnection strategy seemed to be strongly dependent on the patients and their circumstances [6], also requiring a reintubation. Since reintubation may cause serious problems in some patients (and even end with death), many researchers have tried to identify the physiological factors affecting the extubation process. Scientific evidence has shown that the risk of death increases when the patient suffers a failure weaning [1].

Determining the optimal instant of extubation is a nontrivial decission. Though the current reintubation rate is still in the range of 15-30% [1], new indices for accurate prediction are still under investigation. In the last decade, several authors have proposed different methods for data analysis and model inference using only respiratory parameters, such as inspiratory time, expiratory time, breath duration, or tidal volume [7] [8] [9] [10] [11]. Other authors [12] [13] [14] [15] have proposed similar methods extending the use of the afore mentioned parameters with other physiological (age, sex, or blood pressure) biochemical (creatinine, albumin, or hemoglobin), and pathological (such as multiple-organ failure, traumas, or medical scores) data.

Most of the above proposed methods aim to propose an accurate prediction model for successful weaning, while working with a limited data set of observations. Different statistical and data analysis techniques have been used for providing with

useful success indices from the information retrieved from these databases. A linear discriminant and logistic regression classifiers were used to estimate the weaning failure procedure probability [11]. Cluster analysis, together with feature selection algorithms and neural networks, were proposed in [10] in order to classify patients with success or failure in the weaning process. In [13] [9] a predictive model of ventilation was proposed using Support Vector Machines (SVM). These previous works were usually designed from a limited data set, as far as classical statistical and machine learning classification techniques require a homogeneous set of variables to represent a consistent input space for them.

However, a vast amount of the data usually available in ICU are measured and stored in an heterogeneous format. This is caused because time series are usually acquired at different time instants, what represents an heterogeneity in terms of sampling period and number of samples; similar considerations can be done for clinical tests. Missing values, and ocassionally even some incorrect ones, have to be taken into account.

Classical statistics and machine learning techniques usually require of feature extraction and selection preprocessing steps, which are mostly unable to deal with these heterogeneous series. In this setting, we propose the use of two unsupervised statistical learning tools for preprocessing heterogeneous time series, in the sense of non-regular sampling rates and different lengths. The first procedure is the Normalized Compression Distance (NCD) technique [16], used in a number of descriptive and predictive applications [17] [18] [19]. Using the compression length, the NCD technique provides a similarity measure (in terms of entropy) between two sequences, regardless of their sampling frequency and number of samples. In this work, the NCD technique is used to identify patterns of sequences in the weaning process. Then, Multidimensional Scaling (MDS) [20] is applied to locate each sequence in an N-dimensional space, in order to design a subsequent classifier for predicting the result of the weaning process.

In this work, we have benchmarked differents supervised learning methods for weaning prediction [21]. Best performance was provided by Partial Least Squares (PLS) [22], an appropriate alternative to classical methods.

The remaining of the paper is organized as follows. Next section presents the techniques and proposed methodology to estimate the result of the weaning process. Results with real-world weaning data using classical tools and those proposed in this paper are shown in Section III. Conclusions and future work are summarized in Section IV.

II. Methodology and Statistical Methods

The proposed procedure for dealing with the time series associated to w weaning events consists of three stages, graphically represented in Fig. 1. In the first stage, the NCD technique is used to get a dissimilarity measure d between pairs of the w time sequences, $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_w\}$, providing the NCD matrix of size $w \times w$. Then, the MDS technique is used to project the NCD matrix onto an N-dimensional space, getting w points, $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_w\} \in \mathbb{R}^N$. A subsequent classification process is used to distinguish success and failed weanings according to the corresponding targets. Given that conventional classification techniques (such as Neural Networks or SVM) have been described in the weaning sucess prediction literature [10] [13], we present here the PLS technique, a less used technique which has shown to be extremely useful when the number of explanatory variables N exceeds the number of observations w [22] (common scenario in clinical studies with a reduced number of cases). In this work, a theoretical presentation of the NCD, MDS and PLS techniques is complemented with a synthetic example, in order to get a better understanding of the whole methodology shown in Fig. 1.

A. Synthetic Example

Let us consider a binary classification problem, where each class (success and failed) is characterized by a synthetic time pattern of 48 hours, with values in the range [0,1]. Pattern for success is a sinusoid with exponentially decreasing amplitude (see Fig. 2 (a)), and pattern for failure is a triangle (see Fig. 2 (b)).

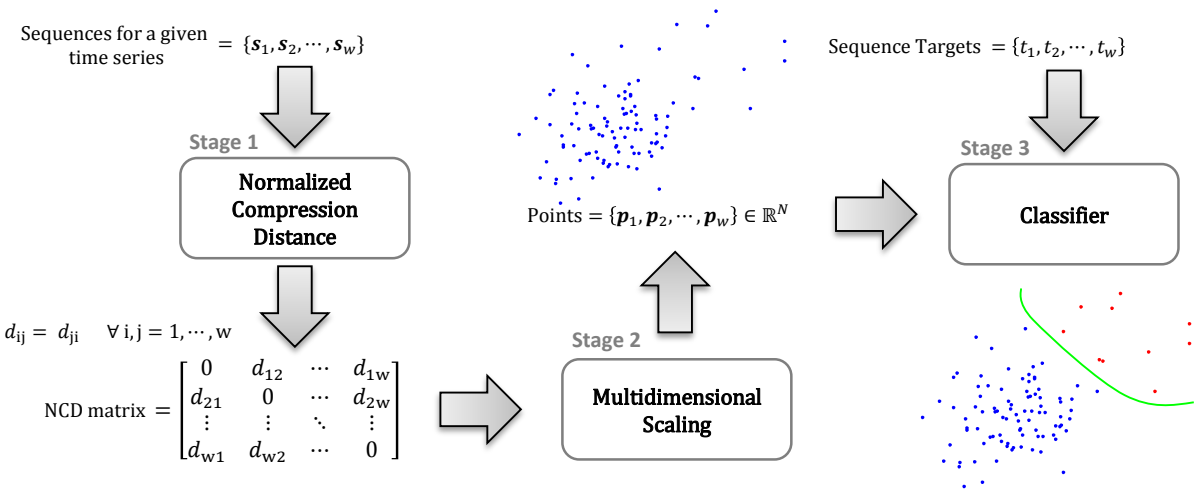


Fig. 1. Diagram of the proposed procedure to deal with heterogeneous time series for failure weaning prediction.

To represent the scenario of our clinical data, both patterns were sampled in a non-regular way (mimimun rate of one second) to provide w = 100 time sequences, $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{100}\}$, corresponding to an imbalanced dataset (10% of sequences were labeled as failed weanings). The number of samples per sequence was a random value between 10 and 150 (typical values in our clinical series). To keep the time reference, the length of every sequence was fixed to 2881 (corresponding to regular sampling of one sample per second for 48 hours). A non-allowed value (e.g. -1) was set in the time instants where no samples were taken. Figures 2 (c) and (d) show one of these sequences for each pattern or class.

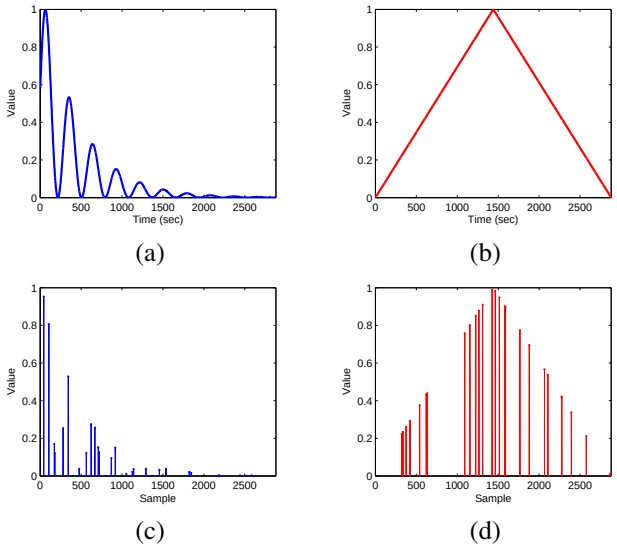


Fig. 2. Time pattern of: (a) success class; (b) failed class; example of sequence for: (c) success class; (d) failed class.

B. Normalized Compression Distance

The aim of the first stage is to obtain a measurement, in terms of distance, for comparing two sequences regardless of their length and sampling frequency. For this purpose we used the NCD technique, closely related to the Kolmogorov complexity. The Kolmogorov complexity of a string $\boldsymbol{s}_i$, $K(\boldsymbol{s}_i)$, is defined as the number of bits of the shortest computer program of the fixed reference computing system capable of producing $\boldsymbol{s}_i$ [23]. Clearly, the choice of the computing system changes the value of $K(\boldsymbol{s}_i)$ in an additive fixed constant. Note that $K(\boldsymbol{s}_i)$ can be considered as the number of bits of the ultimate compressed version of $\boldsymbol{s}_i$ from which $\boldsymbol{s}_i$ can be recovered by a decompression program. Intuitively, $K(\boldsymbol{s}_i)$ corresponds to the minimum amount of information required to generate $\boldsymbol{s}_i$, i.e. to aproximate the entropy of $\boldsymbol{s}_i$.

Given two strings $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, the length of the shortest program computing $\boldsymbol{s}_j$ from $\boldsymbol{s}_i$ is called *information distance*, $E(\boldsymbol{s}_i, \boldsymbol{s}_j)$, and it is defined as [24]:

$$E(\boldsymbol{s}_i, \boldsymbol{s}_j) = K(\boldsymbol{s}_i, \boldsymbol{s}_j) - min\{K(\boldsymbol{s}_i), K(\boldsymbol{s}_j)\} \quad (1)$$

where $K(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the length of the shortest program producing the concatenated pair $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. It was shown in [24] that $E(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is actually a metric and depends on the size of strings. This means that, if for example we consider more than two small strings that differ by an information distance which is large compared to their sizes, then the strings are very different but, if we consider two very large strings that differ by the same (now relatively small) information distance, then they are very similar. Therefore, the information distance itself is not suitable to express true similarity. To solve this problem it is necessary to define a relative distance, what is called Normalized Information Distance (NID), with properties justifying its description as an informative metric [16]:

$$NID(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{K(\boldsymbol{s}_i, \boldsymbol{s}_j) - min\{K(\boldsymbol{s}_i), K(\boldsymbol{s}_j)\}}{max\{K(\boldsymbol{s}_i), K(\boldsymbol{s}_j)\}} \quad (2)$$

NID discovers for every pair of strings the feature in which they are most similar, and expresses that similarity on a scale from 0 to 1. [19].

In the practical use, data compression programs can be applied to approximate the Kolmogorov complexities $K(\boldsymbol{s}_i)$, $K(\boldsymbol{s}_j)$ and $K(\boldsymbol{s}_i, \boldsymbol{s}_j)$. Thus, for a given compressor C, $C(\boldsymbol{s}_i)$ denotes the length, in bits, of the compressed string $\boldsymbol{s}_i$. Using this approximation in (2), we achieve the Normalized Compression Distance (NCD):

$$NCD(\boldsymbol{s}_i, \boldsymbol{s}_j) = d_{ij} = \frac{C(\boldsymbol{s}_i, \boldsymbol{s}_j) - min\{C(\boldsymbol{s}_i), C(\boldsymbol{s}_j)\}}{max\{C(\boldsymbol{s}_i), C(\boldsymbol{s}_j)\}} \quad (3)$$

NCD is a nonnegative number on a scale from 0 to 1 representing how different the two strings are. Low values of NCD represent strings more similar, while values close to correspond to 1 different strings. In practice, NCD values can be higher than 1 for real-world compressors [16].

As it is shown in Fig. 1, the first stage of the proposed procedure provides a symmetric NCD matrix of size $w \times w$. Note that the main diagonal of this matrix has zero values and off-diagonal elements correspond to NCD values between different strings. The gzip compressor has been used in our experiments, though many real-world compressors can be used in practice (zip, bzip2 or PPMZ, among others). Regarding previous synthetic example, time sequences are the strings and the size of the NCD matrix is 100×100.

C. Multidimensional Scaling

The next stage consists of projecting the NCD matrix onto a N-dimensional space by means of the Multidimensional Scaling (MDS) technique (see Fig. 1), also known as Principal Coordinates Analysis. It is an exploratory technique for representing a dissimilarity matrix and visualize the proximity and relationship of the elements in a low-dimensional space. Actually, MDS is a generalization of the idea of Principal Component Analysis (PCA) [20].

Let us consider a symmetric $w \times w$ matrix $\mathbf{M}_{\text{ncd}}$ containing the pairwise dissimilarities of a set of w observations. Thus, d_{ij} represents the dissimilarity between observations i and j (element ij of the NCD matrix). The MDS technique searches

an orthogonal N-dimensional configuration of w points, with $N < w$, $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_w\} \in \mathbb{R}^N$, such that dissimilarities among these points are equal to the dissimilarities provided by the elements of matrix $\mathbf{M}_{\text{ncd}}$. Mathematically, this is equivalent to minimize the following cost function (MDS criterion) [20]:

$$\frac{1}{2}\sum_{i=1}^{w}\sum_{j=1}^{w}(d_{ij} - \parallel \boldsymbol{p}_i - \boldsymbol{p}_j \parallel_2)^2 \tag{4}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. In general it is not possible to find a configuration providing exactly the same dissimilarities. However, approximations can be found (as N increases, approximations are closer to the actual dissimilarities). If w is high, the size of $\mathbf{M}_{\text{ncd}}$ will be also high, and a representation in a low-dimensional space will allow us to understand its structure: proximity between elements, groups, outliers, etc.

When the MDS technique is applied to the NCD matrix of the synthetic example, 100 points of $N = 99$ dimensions were obtained, each point associated to a different sequence. Following with the synthetic example, Fig. 3 represents the two dimensions provided by MDS with highest difference between classes (dimensions 2 and 3). Note that, in this case, classification is easily performed with a linear classifier. However, this is not the typical case with non-synthetic sequences, usually requiring a learning stage to design non-linear classifiers.

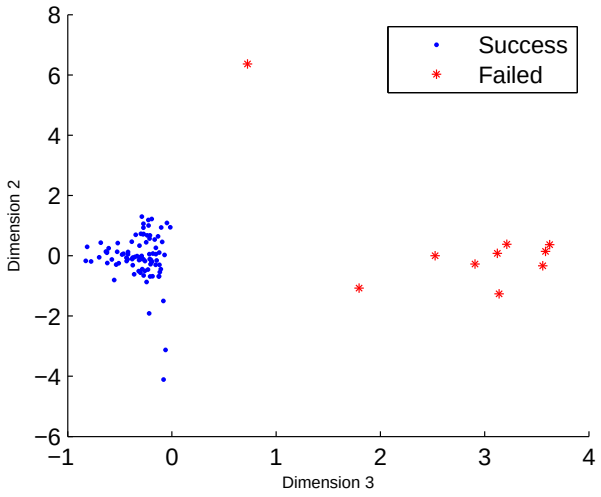


Fig. 3. MDS projection of sequences in synthetic example of Section II-A. Dimensions with the highest difference between classes have been considered.

D. Partial Least Squares

Partial Least Squares (PLS) techniques are used for modeling relations between blocks of variables (e.g., a block of N explanatory variables and another block of M response variables). PLS can be used to tackle regression and classification tasks, as well as dimension reduction and modeling [22]. PLS techniques assume that the observed data are generated by a process driven by a small number of latent (not directly observed) data. PLS extracts orthogonal score vectors (also called latent vectors) by maximising the covariance between blocks of variables; then PLS projects the observed data to its latent structure and use the latent vectors to perform regression of the response variables.

After observing w instances from each block of variables, PLS decomposes the $(w \times N)$ matrix of explanatory variables $\mathbf{P}$ and the $(w \times M)$ matrix of response variables $\mathbf{Y}$ into the form[1]:

$$\begin{aligned}\mathbf{P} &= \mathbf{C}\mathbf{S}^T + \mathbf{R}_P \\ \mathbf{Y} &= \mathbf{L}\mathbf{Q}^T + \mathbf{R}_Y\end{aligned} \tag{5}$$

where $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_v\}$ (components vectors) and $\mathbf{L} = \{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_v\}$ (latent vectors) are $(w \times v)$ score matrices containing the v extracted orthogonal projections of $\mathbf{P}$ and $\mathbf{Y}$, respectively. Matrices of loadings $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_v\}$ and $\mathbf{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v\}$ contain correlations between explanatory variables and component vectors, and between response variables and latent vectors, respectively. $\mathbf{R}_P$ and $\mathbf{R}_Y$ are matrices of residuals. Assuming a linear relation between scores vectors $\boldsymbol{c}$ and $\boldsymbol{\ell}$

$$\mathbf{L} = \mathbf{C}\mathbf{D} + \mathbf{R}_D \tag{6}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{R}_D$ is a matrix of residuals. Replacing (6) in the second equality of (5),

$$\mathbf{Y} = \mathbf{C}\mathbf{D}\mathbf{Q}^T + \mathbf{R}^* \tag{7}$$

where $\mathbf{R}^* = (\mathbf{R}_D\mathbf{Q}^T + \mathbf{R}_Y)$ is a residual matrix. Equation (7) is the decomposition of $\mathbf{Y}$ using ordinary least squares regression with orthogonal vectors $\mathbf{C}$, and reflects the PLS assumption that component vectors are good predictors of $\mathbf{Y}$.

The conventional way to find component and latent vectors is based on the Nonlinear Iterative Partial Leart Squares (NIPALS) algorithm [25], which provides weighting vectors $\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_v\}$ and $\mathbf{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_v\}$, such that:

$$\begin{aligned}cov^2(\boldsymbol{c}, \boldsymbol{\ell}) &= cov^2(\mathbf{P}\boldsymbol{w}, \mathbf{Y}\boldsymbol{u}) \\ &= max_{|\boldsymbol{r}|=|\boldsymbol{b}|=1} cov^2(\mathbf{P}\boldsymbol{r}, \mathbf{Y}\boldsymbol{b})\end{aligned} \tag{8}$$

where $cov(\boldsymbol{c}, \boldsymbol{\ell})$ is the sample covariance between vectors $\boldsymbol{c}$ and $\boldsymbol{\ell}$. Weighting vectors $\boldsymbol{w}$ and $\boldsymbol{u}$ can also be found with algorithms based on eigenvectors descomposition [26], or using other approaches as SIMPLS [27]. Then, from $\boldsymbol{c}$ and $\boldsymbol{\ell}$, the loadings vectors $\boldsymbol{s}$ and $\boldsymbol{q}$ can be computed [22]. Using the relationship [25]:

$$\mathbf{C} = \mathbf{P}\mathbf{W}(\mathbf{S}^T\mathbf{W})^{-1} \tag{9}$$

it is possible to rewrite (7) in terms of the explanatory variables

$$\mathbf{Y} = \mathbf{P}\mathbf{B} + \mathbf{R}^* \tag{10}$$

where $\mathbf{B}$ represents a matrix of regression coefficients

$$\mathbf{B} = \mathbf{W}(\mathbf{S}^T\mathbf{W})^{-1}\mathbf{Q}^T \tag{11}$$

Therefore, linear estimation of $\mathbf{Y}$ is given by

[1] $\mathbf{P}$ and $\mathbf{Y}$ matrices have zero-mean variables.

$$\hat{\mathbf{Y}} = \mathbf{PB} \tag{12}$$

Note that coefficients of **B** denote the influence of each explanatory variable with the response variables. A high value in an entry of **B** indicates that the associated explanatory variable has a high covariance with the associated response variable.

Following with the synthetic example of Section II-A, PLS algorithm is applied to the $w = 100$ data (99 explanatory variables, obtained with the MDS projection, and 1 response variable with the target). Figure 4 represents the 99 regression coefficients of **B** obtained for this example. Note that 2nd and 3rd dimensions are the most influential in the PLS estimation, what is in accordance with the result of Section II-C (dimensions with the highest difference between classes).

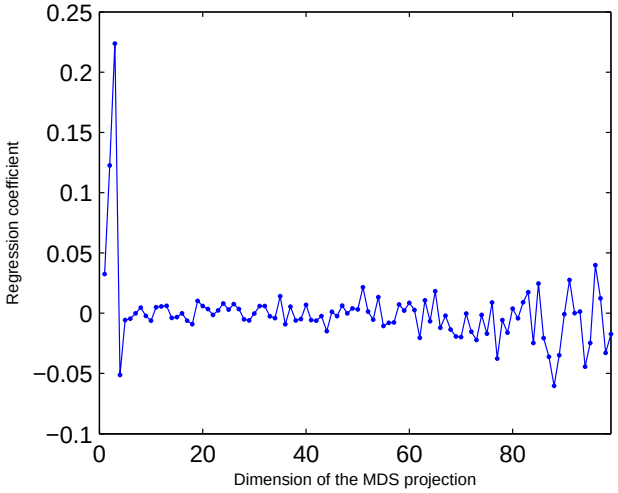


Fig. 4. Regression coefficients in **B** after applying the PLS technique to the result of Stage 2 for the synthetic example in Section II-A.

Since originally PLS is a regression technique and in this work we use it for classification, we established a procedure to fix a threshold and perform a classification task based on $\hat{\mathbf{Y}}$. The threshold value is selected as that with the highest accuracy (percentage of correctly classified weanings) after performing Leave One Out - Cross Validation (later explained in section II-F1).

E. Performance Evaluation

The most common measure for evaluating performance in binary classification problems is the accuracy. However, in cases with imbalanced datasets, a deeper analysis of performance can be provided through the confusion matrix (represented in Table I). For the binary case, the confusion matrix becomes a 2x2 matrix containing the correct classifications in the major diagonal entries and the possible errors in the off-diagonal entries. Each column of the confusion matrix represents the number of instances in a predicted class, while each row represents the number of instances in the actual class. The True Positive (TP) cases refer to the sucess events correctly predicted; True Negative (TN) cases refer to the failed events predicted as failed; False Positive (FP) to the failed events incorrectly classified as SW; and False Negative (FN) to the success events predicted as failed. Using this nomenclature, the following performance measures (focusing in our classification problem, success or failed weaning) can be used [28]:

1) Accuracy: It is the probability of classifying correctly the weanings. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{13}$$

2) Sensitivity: This measure is the probability of classifying correctly the success weanings. It is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

3) Specificity: In contrast to the sensitivity, specificity is the probability of classifying correctly the failed weanings. It is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

4) Balanced Error Rate (BER): This measure weights equally errors of each class. Actually, the BER considers simultaneously the complementaries of the specificity and the sensitivity. It definition is as follows:

$$BER = \frac{1}{2}\left(\frac{FN}{TP + FN} + \frac{FP}{TN + FP}\right) \tag{16}$$

The BER measure is very used with imbalaced dataset. For example, an dataset with 100 instances (90 negative and 10 positive), if all instances are classified as negative,the BER = 0.5, indicating a poor performance regardless of the accuracy obtained (0.9, prior probability of the majority class - baseline accuracy). Note that a lower BER indicates that sensitivity and specificity are high and therefore performance is good.

5) Area Under Curve (AUC): It is the area under the curve obtained by plotting $sensitivity$ against $1 - specificity$ for each decision threshold in the classification.

TABLE I
CONFUSION MATRIX USED IN OUR DOMAIN

Actual \ Predict	Success	Failed
Success	True Positive (TP)	False Negative (FN)
Failed	False Positive (FP)	True Negative (TN)

F. Validation Methods

In this work, two methods based on "resampling" are used for estimating generalization performance.

1) Leave One Out - Cross Validation (LOOCV): This technique [21] partitions the original dataset into w subsets (as many subsets as available instances). A total number of w statistical models are designed, each model being designed on a different combination of $w - 1$ of the w subsets, and performance is evaluated on the partition not used for design (test partition). The model performance is estimated as the

average performance on the w test partitions. LOOCV has been shown to give an almost unbiased estimator of the generalization properties of statistical models, and therefore provides a sensible criterion for model selection and comparison.

2) Bootstrap Resampling: Let us assume a random variable x and a set $\mathbf{X}_w$ of w i.i.d. instances of x. A bootstrap resample is constructed by randomly selecting from $\mathbf{X}_w$ a number of w instances with replacement. This resampling procedure is repeated B times, forming B sets $\mathbf{X}_w^{(b)}$ of w instances each one, $b = 1, \ldots, B$. The bootstrap resamples $\mathbf{X}_w^{(b)}$ are conditionally independent given $\mathbf{X}_w$ and follow the same distribution as x.

Let us assume now that we desire to estimate an statistics θ of x (e.g. mean) using an estimator $\varphi(\cdot)$, where $\hat{\theta}_w = \varphi(\mathbf{X}_w)$ represents an estimation of θ from $\mathbf{X}_w$. If $\varphi(\cdot)$ is applied to the bootstrap resamples $\mathbf{X}_w^{(b)}$, B estimations $\hat{\theta}_w^{(b)}$ are obtained. The consistency of $\hat{\theta}_w$ can be assessed using statistics (such as standard deviation or confidence interval) of the bootstrap estimations. Bootstrap resampling is a robust technique used in scenarios with a reduced number of instances [29], and in this work it is used for assessing the consistency of the performance measurements.

III. Experiments and Results

A. Weaning Data

From January 2010 to December 2011, a total number of 253 patients from ICU of *Hospital Universitario Fundación Alcorcón* under mechanical ventilation were considered. The patient's information was collected according to a protocol approved by the local ethic committee. Selected patients had not suffered a tracheotomy procedure and their mechanical ventilation time was longer than 2 days. Finally, our analysis was performed on 93 intubated patients, providing a total of 104 weaning events which were classified into two classes: 88 events for the SW (success weaning) class and 16 events for the FW (failed weaning) class[2]. In this work, a weaning event corresponds to the FW class when the patient is reintubated within 48 hours after extubation [1].

The weaning dataset were collected using the clinical information system IntelliVue Clinical Information Portfolio (ICIP) [29]. For each event, 20 time series (TS), 15 clinical laboratory parameters (CLP) and 12 general descriptors (GD) (if available) were collected at least once during 48 hours previous to the weaning event (see Table II). From a clinical point of view, these variables are potentially influential in the weaning event. Initially, ICIP did not allow automatical extraction of variables. A manual extraction of the data would be a very long and tedious task (for months) with error probability high. For helping in this work, Philips developed a tool that facilitated this process by reducing a high degree of extraction time (two days). Even so, some of the GD variables had to be collected manually. A remark is interesting here, being necessary to assign a value to these variables in order to consider them in the analysis.

TS variables were sampled in a non-regular manner and at a minimum rate of one second, providing a number of values per variable fluctuating between 1 and 138. Since these variables have a time stamp per value, we considered them as time sequences. Ramsay and Glasgow scales variables were discarded in this work because they were not available for many of the events (41%). Fig. 5 shows an example of two TS variables for two weaning events (SW and FW classes) during 48 hours before the weaning event (2881 seconds, a sample per second where sample 0 corresponds to the 48 hours before the weaning event, and that the sample 2881 corresponds to the previous second to event). Note that, in contrast to the synthetic example of Section II-A, it is not evident to devise a characteristic pattern for each class.

[2]Note that in this case dataset are imbalanced.

TABLE II
Time Series (TS), Clinical Laboratory Parameters (CLP) and General Descriptors (GD) for each weaning event

TS	CLP	GD
Heart rate	Albumin	APACHE3
Diastolic blood pressure	creatinine	SAPS2
Systolic blood pressure	Hematocrit	SAPS3
Temperature	Hemoglobin	SOFA1
Glasgow scale	Leukocytes	SOFA2
Ramsay scale	C Reactive Protein	% IPPV
Respiratory rate	SBC	% BIPAP
Resistance	Urea	% ASB
Peep	Arterial pCO_2	% O_2TT
Support pressure	Venous pCO_2	Time MV
Mean pressure	Arterial pH	Age
Plateau pressure	Venous pH	Sex
Peak pressure	Arterial pO_2	
Inspiratory time	Lactic Acid	
Compliance	Procalcitonin	
Inspiratory flow		
Expired minute volume		
Tidal volume		
spO_2		
fiO_2		

In the CLP group, each variable had a reduced number of values (from 0 to 7, depending on the weaning event), and the mean value was computed as the representative value; variables with no values (missing data) were imputed to zero. Regarding the group of GD variables, they can be numerical or categorical (for example sex or APACHE3) and just have one value per variable.

B. Experiments with conventional tools

Several experiments were performed, some of them to reproduce the schemes proposed in other studies. The first round of experiments considered features of the three kinds of variables (TS, CLP and GD). Thus, eight statistics were obtained from each TS variable: minimum, maximum, standard deviation, variance, interquartile range, mean, median, and summatory. These statistics, together with the CLP and GD variables, provided a total of 187 features (8 statistics x 18 TS + 15 CLP + 12 GD). Baseline accuracy was the best result.

Since the number of features was larger than the number of instances, two feature selection procedures were applied to select the most relevant features and check if better performance was possible. The first procedure was the Mann-Whitney test [30], applied to each feature to test its significance to distinguish between FW and SW classes; features features with

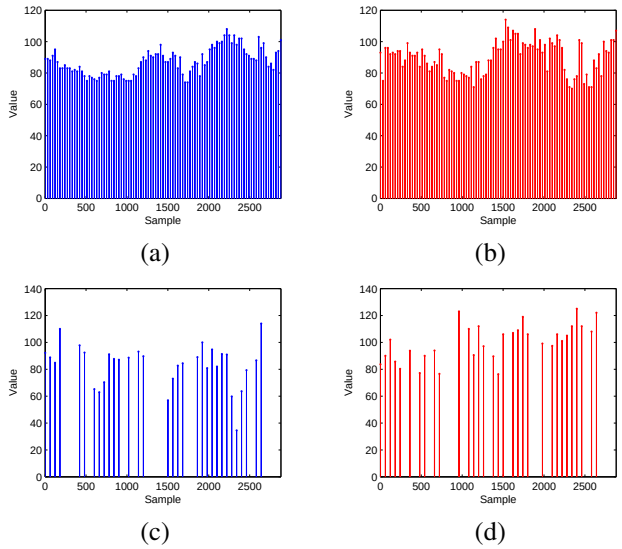


Fig. 5. Time sequences for two TS variables: heart rate (upper panels) and expired minute volume (lower panels). Left panels correspond to the SW class, and right ones to the FW class.

a p-value lower than 0.1 were selected. The second selection procedure, proposed in [31], is a backward method based on bootstrap and SVM classifiers. Selected features with each procedure were used to design three classifiers: linear SVM, nonlinear SVM with RBF kernel, and k-NN. Baseline accuracy was not improved in any case. LOOCV was applied both in the feature selection process and in the classifier design.

In the second round of experiments, each type of variable (TS, CLP and GD) was considered independently for predicting the weaning event: 144 features for TS, 15 features for GLP and 12 for GD variables. Experiments with TS variables were completed extracting just one feature per time sequence: (1) median, and (2) trend (obtained as the slope of the least squares regression line). Note that three experiments with different features were performed with the TS variables. Classification with each group of variables just improved the baseline accuracy when the 144 features from TS variables were considered, providing an accuracy of 85.57%.

To select the most relevant features of each group of variables, filter, wrapper and embedded feature selection approaches available in the Weka tool [32] were applied. Filter methods do not involve any classifier in the selection process (i.e, selection and classification are independent processes); wrapper methods use the performance of certain classifier as a criterion for feature selection; and embedded methods perform simultaneously both selection and classification processes. In this work we considered the C4.5 algorithm as the embedded approach. Selected features were used for weaning event prediction with decision trees, MLP, SVM, and the Adaboost M1 method with the aforementioned base classifiers.

Regarding experiments with TS variables, the feature selection procedure did not improve the accuracy when the trend was considered. With the median features, best accuracy of 87.5% was achieved with the most selected TS variables according to the aforementioned procedures (heart rate, systolic blood pressure, respiratory rate, peep, mean pressure, plateau pressure, compliance, inspiratory flow and fio2). With the set of 8 statistics per TS variable, accuracy of 88.5% was obtained with just 4 features (statistics of interquartile range and mean, and variables heart rate, compliance, and systolic blood pressure), selected with the wrapper approach.

In the third round of experiments, features obtained from the MDS projection of each TS variable were used for classification. Two selection procedures were considered here. In the first one, the Mann-Whitney test was applied to select, from each MDS projection, the feature with the lowest p-value. Best accuracy of 92.3% was obtained with the classifier designed with the most relevant selected projections: inspiratory time and systolic blood pressure variables (subset of features highly correlated with the class and with low correlation among them).

From the experiments presented in this section, it is clear that trying to predict the outcome of the weaning event from heterogeneous data is not a simple task. Furthermore, in the above experiments, temporal reference in TS variables was not taken into consideration. In order to deal with the raw data while maintaining the temporal reference, an investigation was made to deal with heterogeneous time series, leading to the procedure proposed in Section II.

C. Experiment with the proposed procedure

The NCD technique described in Section II-B was applied to the 18 TS variables mentioned in Section III-A. The length of each sequence was set to 2881 (sampling frequency of 1 Hz for 48 hours before the weaning event). A total of 18 NCD matrices of size 104×104 were obtained, one for each TS variable. Then, the MDS technique was applied to each NCD matrix. As a result, 18 matrices of size $104 \times N$ were obtained, with N potentially different for each TS variable. In our experiments, N was in the range [47,98].

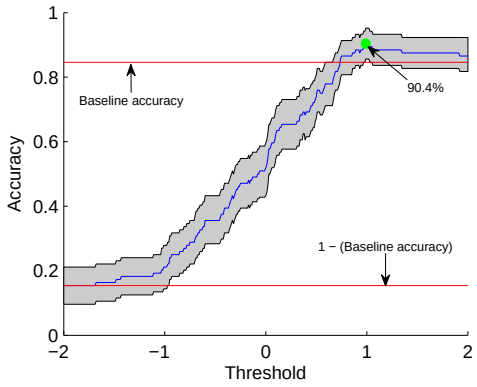
Three classifiers (linear SVM, nonlinear SVM with RBF kernel, and k-NN) were designed for classifying weaning events considering just one TS variable. The N features obtained from the MDS technique were used for this purpose. Accuracy with LOOCV did not exceed the baseline accuracy in any case. To check if another classification technique worked better, PLS was applied (decision threshold was chosen as that providing the highest accuracy). Table III shows the acurracy, BER and AUC obtained with LOOCV when the PLS technique was applied to the MDS projections of each TS variable. The best result was obtained with the diastolic blood preassure variable: accuracy of 90.4% and BER of 28.7%.

Accuracy and BER obtained with PLS and different decision thresholds for the diastolic blood pressure variable are shown in Figure 6. For assessing the consistency of the performance measurements, bootstrap resampling was applied to the predictions provided by the PLS technique. The gray area represents the bootstrap confidence interval of 95% for each threshold, while the blue lines represents the median values. The best performance (accuracy of 90.4% and BER 28.7%, green points in Figs. 6 (a) and (b)) is obtained for a threshold of 1 (confusion matrix in Table I). Note that the number of false positives is high (9 instances), while that of false negatives is

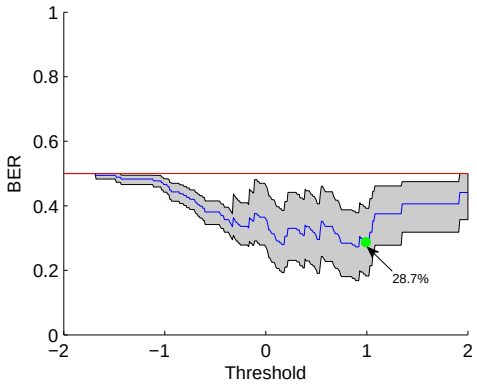
TABLE III
LOOCV RESULTS FOR THE PLS MODEL USING THE MDS PROJECTION OF EACH TS VARIABLE

Time serie	Accuracy %	BER %	AUC %
Diastolic blood pressure	90.4	28.7	78.9
Peak pressure	86.5	43.8	61.5
fiO_2	86.5	43.8	55.5
Systolic blood pressure	86.5	43.4	52.3
Resistance	85.6	46.9	63.6
Inspiratory flow	85.6	46.7	53.2
Peep	85.6	46.9	53.0
spO_2	85.6	41.8	51.9
Respiratory rate	84.6	50.0	66.8
Support pressure	84.6	50.0	64.0
Plateau pressure	84.6	50.0	59.0
Inspiratory time	84.6	47.4	56.3
Compliance	84.6	50.0	56.3
Mean pressure	84.6	50.0	53.8
Temperature	84.6	50.0	52.8
Expired minute volume	84.6	50.0	51.1
Heart rate	83.7	50.0	54.3
Tidal volume	80.8	52.3	63.1

low (1 instance). This fact is produced because the SW class has a stronger influence on the PLS estimation (imbalaced dataset), making it difficult classification of the FW events.



(a)



(b)

Fig. 6. Diastolic blood pressure variable. Accuracy (a) and BER (b) obtained with each decision threshold using the PLS technique. Gray area represents the bootstrap 95% confidence interval for the performance measurements. Blue line corresponds to the median values.

Figure 7 represents the PLS regression coefficients for the $N = 94$ features with the diastolic blood pressure variable. In

TABLE IV
CONFUSION MATRIX FOR A THRESHOLD THE BEST PERFORMANCE FOR THE DIASTOLIC BLOOD PRESSURE VARIABLE

Actual \ Predict	Success	Failed
Success	87	1
Failed	9	7
Accuracy = 90.4 % **AUC** = 78.9 % **BER** = 28.7 %		

contrast to the results of the synthetic example of Section II-A, where one dimension highlighted, now several dimensions are influential in the PLS estimation (regression coefficients are different to 0 -red line-). This result makes it difficult an interpretation of what are the main influential dimensions and what not to differentiate the classes, which indicates that the variable information completely (and not just one dimension) allows to perform this task.

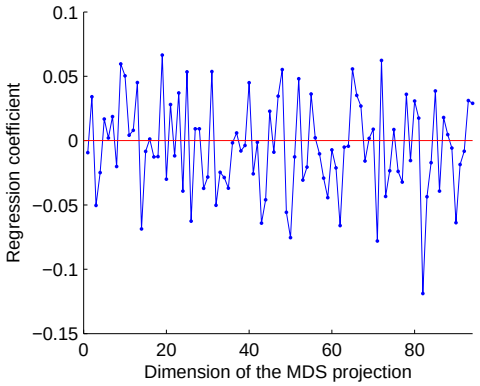


Fig. 7. Regression coefficients after applying the PLS technique to the MDS projections of diastolic blood pressure variable.

IV. CONCLUSION

We presented in this work a procedure to deal with heterogeneous time series for failure weaning prediction. Our procedure involves three stages: (1) obtention of the similarity matrix between pairs of sequences using the NCD technique; (2) projection of these distances on an N-dimensional space through the MDS technique; and (3) obtention of the projections maximizing the percentage of covariance between explanatory variables and targets, using the PLS technique for predicting the weaning outcome.

Application of the proposed procedure to the synthetic example of Section II-A shows its capability to deal with time series regardless of the sampling frequency and the number of samples. The joint use of the two statistical techniques (NCD and MDS) allows us to provide a compact input space to design a statistical classifier.

In our experiments with real-world weaning data, 18 TS variables have been used to validate the proposed procedure. Only one of them, diastolic blood preassure, reached significant results when it was considered without combination to other variables, achieving with PLS 90.4% and 28.7% of accuracy and BER, respectively. For comparison with other

methods, several experiments were performed using feature selection approaches and other classification techniques such as decision trees, MLP, SVM, k-NN and Adaboost. The results of these experiments guided the research towards the joint use of TS variables.

A number of variables for prediction of weaning events have been analyzed using statistical methods. Though this work has been focused on time series, we propose to develop an extension to adapt the proposed procedure and deal with other type of variables (not time series), including complementary information for improving global performance.

V. Acknowledgement

The authors would like to thank María Tato Cerdeiras, specialist of health care applications in Philips Ibérica, Madrid, Spain, for her help in the weaning data extraction task. José María Lillo would also like to thank PhD. Ricardo Santiago Mozos for his support in the understanding of the PLS technique (Section II-D).

This work has been partly supported by project TEC2010-19263.

References

[1] M. J. Tobin, *Principles and Practice of Mechanical Ventilation*, 2nd ed. McGraw-Hill, 2006.

[2] N. R. MacIntyre, "Evidence-based ventilator and discontinuation," *Respiratory Care*, vol. 49, no. 7, pp. 830–836, 2004.

[3] ——, "Evidence-based guidelines for weaning and discontinuing ventilatory support," *Chest*, vol. 120, no. 6, pp. 375–395, 2001.

[4] B. Blackwood, F. Alderdice, K. Burns, C. Cardwell, G. Lavery, and P. O'Halloran, "Use of weaning protocols for reducing duration of mechanical ventilation in critically ill adult patients: Cochrane systematic review and meta-analysis," *BMJ*, vol. 342, pp. 1 – 14, 2011.

[5] D. Scheinhorn, D. Chao, M. Stearn-Hassenpflug, and W. Wallace, "Outcomes in post-icu mechanical ventilation a therapist-implemented weaning protocol," *Chest*, vol. 119, no. 1, pp. 236–242, 2001.

[6] A. Bruton, J. Conway, and S. Holgate, "Weaning adults from mechanical ventilation," *Physiotherapy*, vol. 85, no. 12, pp. 652–661, 1999.

[7] P. C. de-la Higuera, M. Martín-Fernández, and C. Alberola-López, "Weaning from mechanical ventilation: A retrospective analysis leading to a multimodal perspective," *IEEE Trans. Biomedical Engineering*, vol. 53, no. 7, pp. 1330–1345, 2006.

[8] P. C. de-la Higuera, F. Simmross-Wattenberg, M. Martín-Fernández, and C. Alberola-López, "A multichannel model-based methodology for extubation readiness decision of patients on weaning trials," *IEEE Trans. Biomedical Engineering*, vol. 56, no. 7, pp. 1849–1863, 2009.

[9] B. Giraldo, A. Garde, C. Arizmendi, R. Jané, S. Benito, I. Díaz, and D. Ballesteros, "Support vector machine classification applied on weaning trials patients," in *EMBS Annual International Conference*, New York City, USA, 30 August - 3 September 2006, pp. 5587–5590.

[10] C. Arizmendi, E. Romero, R. Alquezar, P. Caminal, I. Diaz, S. Benito, and B. Giraldo, "Data mining of patients on weaning trials from mechanical ventilation using cluster analysis and neural networks," in *EMBS Annual International Conference*, Minneapolis, Minnesota, USA, 3-6 September 2009, pp. 4343–4346.

[11] J. Preciado and B. Giraldo, "Análisis y clasificación del patrón respiratorio de pacientes en proceso de retirada del ventilador mecánico," *Revista Ingeniería Biomédica*, vol. 5, no. 9, pp. 43–49, 2011.

[12] H. Jiin-Chyr, C. Yung-Fu, L. Hsuan-Hung, C.-H. L., and J. Xiaoyi, "Construction of prediction module for successful ventilator weaning," in *New Trends in Applied Artificial Intelligence*, ser. Lecture Notes in Computer Science, H. Okuno and M. Ali, Eds. Springer Berlin / Heidelberg, 2007, vol. 4570, pp. 766–775.

[13] Y. Hao-Yung, H. Jiin-Chyr, C. Yung-Fu, J. Xiaoyi, and C. Tainsong, "Using support vector machine to construct a predictive model for clinical decision-making of ventilation weaning," in *International Joint Conference on Neural Networks*, Hong Kong, China, 1-8 June 2008, pp. 3981–3986.

[14] Y.-F. Chen, Y.-F. Huang, X. Jiang, Y.-N. Hsu, and H.-H. Lin, "Design of clinical support systems using integrated genetic algorithm and support vector machine," in *Computer Analysis of Images and Patterns*, X. Jiang and N. Petkov, Eds. Springer Berlin / Heidelberg, 2009, vol. 5702, pp. 791–798.

[15] S. Burns, C. Fisher, S. Tribble, R. Lewis, P. Merrel, M. Conaway, and T. Bleck, "The relationship of 26 clinical factors to weaning outcome," *American Journal of Critical Care*, vol. 21, no. 1, pp. 52–58, 2012.

[16] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, "The similarity metric," *IEEE Trans. Information Theory*, vol. 50, no. 12, pp. 3250 – 3264, 2004.

[17] A. J. Pinho and P. J. S. G. Ferreira, "Image similarity using the normalized compression distance based on finite context models," in *International Conference on Image Processing*, 2011, pp. 1993 – 1996.

[18] S. Axelsson, "Using normalized compression distance for classifying file fragments," in *Availability, Reliability, and Security*, 2010, pp. 641 –646.

[19] R. L. Cilibrasi and P. M. Vitányi, "The google similarity distance," *IEEE Trans. Knowledge and Data Endineering*, vol. 19, no. 3, pp. 370 –383, 2007.

[20] D. Peña, *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[22] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection Techniques, Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer - Verlag, 2006, pp. 34 – 51.

[23] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, 3rd ed. Springer, 2008.

[24] C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. H. Zurek, "Information distance," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1407 – 1423, 1998.

[25] H. Wold, "Nonlinear estimation by iterative least squares procedures," in *Researh Papers in Statistics*, F. David, Ed. Wiley, New York, 1966, pp. 411 – 444.

[26] A. Höskuldsson, "Pls regression methods," *Journal of Chemometrics*, vol. 2, pp. 211 – 228, 1988.

[27] S. de Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, pp. 251 – 263, 1998.

[28] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations And Applications*. Springer-Verlag, 2006.

[29] (2012) Intellivue clinical information portfolio - philips. [Online]. Available: http://www.healthcare.philips.com/es_es/products/patient_monitoring/products/icip/

[30] J. D. Gibbons and S. Chakraborti, *Gibbons, J. D. Nonparametric Statistical Inference.*, 4th ed. Marcel Dekker, 1985.

[31] F. Alonso-Atienza, J. L. Rojo-Álvarez, A. Rosado-Muñoz, J. J. Vinagre, A. García-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1956 – 1967, 2012.

[32] (2012) Weka 3 - data mining with open source machine learning software in java. [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/