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

J.M. Lillo-Castellano[a,*], I. Mora-Jiménez[a], R. Santiago-Mozos[a], J.L. Rojo-Álvarez[a], J. Ramiro-Bargueño[a], A. Algora-Weber[b]

[a] *Signal Theory and Communications Department, Rey Juan Carlos University, Camino del Molino s/n, 28943, Fuenlabrada, Madrid, Spain*
[b] *Intensive Care Unit, Alcorcón Foundation University Hospital, C/Budapest 1, 28922, Alcorcón, Madrid, Spain*

Abstract

In the Intensive Care Unit of a hospital (ICU), weaning can be defined as the process of gradual reduction in the level of mechanical ventilation support. A failed weaning increases the risk of death in prolonged mechanical ventilation patients. Different methods for weaning outcome prediction have been proposed using variables and time series extracted from the monitoring systems, however, monitored data are often non-regularly sampled, hence limiting its use in conventional automatic prediction systems. In this work, we propose the joint use of two statistical techniques, Normalized Compression Distance (NCD) and Multidimensional Scaling (MDS), to deal with data heterogeneity in monitoring systems for weaning outcome prediction. A total of 104 weanings were selected from 93 patients under mechanical ventilation from the ICU of Hospital Universitario Fundación Alcorcón; for each weaning, time series (TS), clinical laboratory and general descriptors variables were collected during 48 hours previous to the moment of withdrawal mechanical support (extubation). The TS diastolic blood pressure variable provided the best weaning prediction, with an

*Corresponding author
Email addresses: josemaria.lillo@urjc.es (J.M. Lillo-Castellano), inmaculada.mora@urjc.es (I. Mora-Jiménez), ricardo.santiago.mozos@urjc.es (R. Santiago-Mozos), joseluis.rojo@urjc.es (J.L. Rojo-Álvarez), julio.ramiro@urjc.es (J. Ramiro-Bargueño), aalgora@fhalcorcon.es (A. Algora-Weber)

improvement of 37% in the error rate regarding the physician decision. This result shows that the joint use of the NCD and MDS efficiently discriminates heterogeneous time series.

Keywords: Weaning, Extubation, Intensive Care Unit, Normalized Compression Distance, Multidimensional Scaling, Partial Least Squares.

1. Introduction

In daily routine of a hospital Intensive Care Unit (ICU), patients are often assisted with mechanical ventilation, which 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 consists in a gradual removal of the mechanical respiratory support (Tobin, 2006). Although current mechanical ventilators are sophisticated devices 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 patient distress, causing difficulty in reestablishing artificial airways and compromising gas exchange (MacIntyre, 2004). 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 hospital economic cost (MacIntyre, 2001). Hence, two main questions have to be taken into account in the weaning setting, specifically, how can the physician decide the best extubation instant, and which information can be used to support this decision.

Nowadays, physicians use their knowledge and own experience to start the patient weaning and select the most appropriate procedure (Blackwood et al., 2011). Currently, the most used method for weaning consist in assessing the patient's respiratory status by observing either his spontaneous breathing through a T-Tube circuit (T-Tube Test) or his breathing while assisted by a low pressure support. If the patient tolerates the test and the physician considers the

weaning is appropriate, the patient is extubated. An example of an alternative method is presented in (Scheinhorn et al., 2001), where a therapist-implemented protocol was used to extubate patients from prolonged mechanical ventilation for reducing the weaning duration. However, the disconnection strategy seemed to be strongly dependent on the patients and their circumstances (Bruton et al., 1999), and it often required a reintubation. Since reintubation may cause serious problems and even exitus, many researchers have tried to identify the physiological factors affecting the weaning. Scientific evidence has shown that the risk of death increases when the patient suffers a failed weaning (Tobin, 2006).

To determine the extubation optimal instant is a nontrivial decision. The current reintubation rate is still in the range of 15-30% and indices for extubation instant prediction are still under active investigation (Tobin, 2006). In the last decade, several authors have proposed different methods for data analysis and model inference using only respiratory parameters, such as inspiratory and expiratory time, breath duration, or tidal volume (Casaseca-de-la Higuera et al., 2006; Giraldo et al., 2006; Arizmendi et al., 2009; Casaseca-de-la Higuera et al., 2009; Preciado and Giraldo, 2011). Other authors have proposed similar methods combining the aforementioned parameters with other physiological (age, sex, or blood pressure), biochemical (creatinine, albumin, or hemoglobin) (Burns et al., 2012), and pathological (such as multiple-organ failure, traumas, or medical scores) data (Jiin-Chyr et al., 2007; Hao-Yung et al., 2008; Yung-Fu et al., 2009).

Most of the previous methods propose a prediction model for weaning outcome, working with a limited number of cases and an homogeneous set of variables. Giraldo et al. (2006) and Hao-Yung et al. (2008) employ Support Vector Machines (SVM) for constructing a predictive model of weaning outcome. Arizmendi et al. (2009) propose cluster analysis together with feature selection algorithms and neural networks to determine the weaning outcome. Preciado and Giraldo (2011) use a linear discriminant and logistic regression to estimate the probability of failed weaning.

In ICU, a vast amount of data are usually measured and stored in an hetero-

geneous way: 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. In addition, missing and occasionally incorrect values have to be dealt with.

Classical statistics and machine learning techniques usually require feature extraction and selection stages, which are mostly unable to deal with heterogeneous time series. In this setting, we propose the joint use of two unsupervised statistical learning tools: Normalized Compression Distance (NCD) (Li et al., 2004) and Multidimensional Scaling (MDS) (Jolliffe, 2002). The NCD technique comes from Information Theory and has been successfully applied to a number of descriptive and predictive applications (Cilibrasi and Vitányi, 2007; Axelsson, 2010; Pinho and Ferreira, 2011). By using the compression length, the NCD technique provides a similarity measure between two sequences (in terms of their mutual information), regardless of their sampling frequency and number of samples. In this work, the NCD technique is used to identify patterns in the time series of the weaning variables. MDS is applied to locate each sequence as a point in an N-dimensional space, which is the input of a subsequent classifier for predicting the weaning outcome. In this work different classifiers have been benchmarked for this purpose. Best performance was provided by Partial Least Squares (PLS) (Rosipal and Krämer, 2006).

The remaining of the paper is organized as follows. Next section presents the techniques and proposed methodology to predict the weaning outcome from heterogeneous time series variables. Results with real-world data using classical tools and those proposed in this paper are shown in Section 3. Conclusions are presented in Section 4.

2. Methodology and statistical methods

Let us consider a given set of w labeled time sequences $\{(\boldsymbol{s}_i, t_i)\}_{i=1}^{w}$, with $\boldsymbol{s}_i$ being the i-th sequence of a time series weaning variable, and t_i its associated label $\{failure, success\}$. The aim is to infer a weaning outcome prediction

model from the w labeled time sequences. For this purpose, a procedure of three stages (graphically represented in Fig. 1) has been proposed:

Stage 1: The NCD technique is applied to the set of w sequences $\{\boldsymbol{s}_i\}_{i=1}^{w}$. A matrix of size $w \times w$ (the named NCD matrix) is obtained, whose elements are a dissimilarity measure d_{ij} between pairs of sequences $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

Stage 2: The NCD matrix is projected onto an N-dimensional space by applying the MDS technique. The result of this stage is a set of points $\{\boldsymbol{p}_i\}_{i=1}^{w} \in \mathbb{R}^N$, each point associated to a different sequence.

Stage 3: The points yielded in Stage 2 are used, together with labels $\{t_i\}_{i=1}^{w}$ of original sequences, to design a classifier to distinguish between successful and failed weanings.

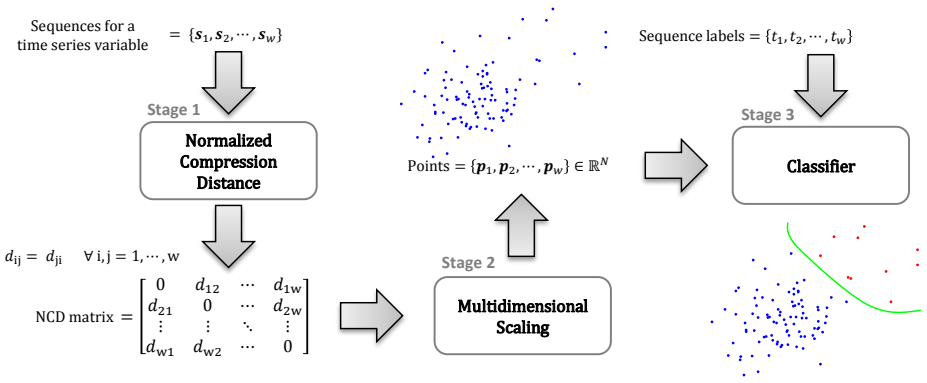


Figure 1: Diagram of the proposed procedure to deal with heterogeneous time series for weaning outcome prediction.

Since conventional classification techniques (such as Neural Networks or SVM) have been described in the weaning outcome prediction literature (Hao-Yung et al., 2008; Arizmendi et al., 2009), we present here the PLS technique (Rosipal and Krämer, 2006), which has been shown extremely useful when the number of explanatory variables N exceeds the number of instances w (a common scenario in clinical studies with a reduced number of instances). In this work, a description of the NCD, MDS and PLS techniques is complemented

with a synthetic example, in order to get a better understanding of the proposed methodology.

2.1. Normalized Compression Distance

The aim of the first stage is to obtain a measure, in terms of distance, for comparing two sequences regardless of their number of elements and sampling frequency. For this purpose we use the NCD technique, closely related to the Kolmogorov complexity. The Kolmogorov complexity of a string $\boldsymbol{s}_i$, named $K(\boldsymbol{s}_i)$, is defined as the number of bits of the shortest computer program of the fixed reference computing system capable of producing $\boldsymbol{s}_i$ (Li and Vitányi, 2008). Hence, $K(\boldsymbol{s}_i)$ 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 a quantity approximately equal to the entropy of the entity that generated $\boldsymbol{s}_i$ multiplied by the length 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 (Bennett et al., 1998) as:

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

where $K(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the length of the shortest program producing the concatenated pair $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Bennett et al. (1998) have shown that $E(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is actually a metric and depends on the length of the strings. For example, if the information distance q between two short strings is large in comparison to their lengths, then the strings are very different; but if two long strings have the same value q for the information distance, since now q is small compared to the strings lengths, then those strings are very similar. Therefore, the information distance itself is not suitable to express true similarity. To solve this problem, Li et al. (2004) defined a relative measure called Normalized Information Distance (NID):

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

NID expresses the similarity between every pair of strings on a scale from zero to one (Cilibrasi and Vitányi, 2007).

In the practical use, data compressors 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 version of the string $\boldsymbol{s}_i$. Using this approximation in (2), the Normalized Compression Distance (NCD) is achieved:

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

which is a non negative number on a scale from zero to one: values of NCD close to zero represent similar strings, while values close to one correspond to different strings. In practice, NCD values can be slightly higher than 1 for real-world compressors (Li et al., 2004).

The first stage of the proposed procedure provides us with an almost symmetric NCD matrix (of size $w \times w$) with entries almost null in the main diagonal. In practice, we force the NCD matrix to be symmetric and have zero values in the main diagonal (see Fig. 1). The *gzip* compressor has been used in our experiments, though other real-world compressors can be used (e.g. *zip*, *bzip2*, *LZMA* or *PPMZ*).

2.2. Multidimensional Scaling

The second stage takes the NCD matrix an projects it onto a N-dimensional space using the Multidimensional Scaling (MDS) technique (see Fig. 1), also known as Principal Coordinates Analysis (Jolliffe, 2002). This is an exploratory technique for representing a dissimilarity matrix and visualizing the proximity of the sequences in a low-dimensional space.

Let us consider the symmetric matrix $\mathbf{M}_{\text{ncd}}$ containing the pairwise dissimilarities of a set of w instances. The MDS technique searches an orthogonal N-dimensional configuration of w points, $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_w\} \in \mathbb{R}^N$ $(N < w)$, such that dissimilarities among these points are as close as possible 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) (Jolliffe, 2002):

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

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). A representation in a low-dimensional space will allow us to understand data structure, for instance, proximity between sequences, groups or outliers.

2.3. 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), as well as for dimension reduction (Rosipal and Krämer, 2006). PLS techniques assume that the observed data are generated by a process driven by a small number of latent (not directly observed) components. PLS extracts orthogonal[1] latent vectors (also called score vectors) by maximizing the covariance between blocks of variables; then PLS projects the observed data (MDS points in our case) to its latent structure and use the latent vectors to perform regression of the response variables.

PLS decomposes the zero-mean $(w \times N)$ matrix of explanatory variables $\mathbf{P}$ and the the zero-mean $(w \times M)$ matrix of response variables $\mathbf{Y}$ into the form:

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

where $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_v\}$ and $\mathbf{L} = \{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_v\}$ are $(w \times v)$ latent matrices containing the v extracted latent vectors of $\mathbf{P}$ and $\mathbf{Y}$ respectively, and $\mathbf{R}_P$ and $\mathbf{R}_Y$ are matrices of residuals. Loading matrices $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_v\}$ and

[1] This orthogonality avoids numerical problems that may arise in Ordinary Least Squares when the variables are highly colinear.

$\mathbf{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v\}$ contain correlations between $\mathbf{P}$ and $\mathbf{C}$; and between $\mathbf{Y}$ and $\mathbf{L}$, respectively. Assuming a linear relationship between latent vectors $\boldsymbol{c}$ and $\boldsymbol{\ell}$, it is possible to express $\mathbf{L}$ as:

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

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

$$\mathbf{Y} = \mathbf{CDQ}^T + \mathbf{R}^* \tag{7}$$

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 with orthogonal vectors $\mathbf{C}$, and reflects the assumption that latent vectors of $\mathbf{P}$ are good predictors of $\mathbf{Y}$.

The conventional way to find latent vectors is based on the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Wold, 1966), which provides weighting matrices $\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_v\}$ and $\mathbf{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_v\}$ such that:

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

where $\text{cov}(\boldsymbol{c}, \boldsymbol{\ell})$ 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 eigenvector decomposition (Höskuldsson, 1988), or using other approaches as SIMPLS (Jong, 1998). After the extraction of the score vectors $\boldsymbol{c}$ and $\boldsymbol{\ell}$, the loading vectors $\boldsymbol{s}$ and $\boldsymbol{q}$ can be computed as coefficients of regressing $\mathbf{P}$ on $\boldsymbol{c}$ and $\mathbf{Y}$ on $\boldsymbol{\ell}$, respectively (Rosipal and Krämer, 2006). Using the relationship $\mathbf{C} = \mathbf{PW}(\mathbf{S}^T\mathbf{W})^{-1}$ (Wold, 1966), it is possible to rewrite (7) in terms of the explanatory variables

$$\mathbf{Y} = \mathbf{PB} + \mathbf{R}^* \tag{9}$$

where $\mathbf{B} = \mathbf{W}(\mathbf{S}^T\mathbf{W})^{-1}\mathbf{Q}^T$ is a regression coefficients matrix. Therefore, linear estimation of $\mathbf{Y}$ is given by:

$$\hat{\mathbf{Y}} = \mathbf{PB} \tag{10}$$

Note that values of $\mathbf{B}$ denote the influence of each explanatory variable on the response variables. A high absolute value in an entry of $\mathbf{B}$ indicates that

the associated explanatory variable has a high covariance with the associated response variable.

Since originally PLS is a regression technique and the weaning problem has been defined as a classification task, a procedure for thresholding $\hat{\mathbf{Y}}$ has been established. The threshold is selected as the one providing the highest accuracy (percentage of correctly classified weanings) after performing the two resampling methods revised in Section 2.5.

2.4. Synthetic Example

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

To represent a similar scenario to that of our clinical data, both time series patterns were sampled in a non-regular way (minimum rate of one minute) to provide $w = 100$ time sequences, corresponding to an imbalanced dataset (10% of sequences were labeled as failure). The number of samples per sequence was a random value between 10 and 150 (typical values in our clinical series, see an example in Figs. 2c and 2d).

The NCD technique was applied to the sequences artificially generated, and a NCD matrix of size 100×100 was obtained (*gzip* compressor was used in this case). Subsequently, the MDS technique was applied to the NCD matrix and 100 points of $N = 99$ dimensions were obtained (each point associated to a different sequence). Finally, the PLS algorithm was applied for prediction outcome (99 explanatory variables and 1 response variable), and the corresponding regression coefficients are shown in Fig. 3.

For visualization purposes, in Fig. 4 the MDS points are depicted on a plane using the most influential dimensions according to the PLS algorithm (2nd and 3rd dimensions, see Fig. 3). Though in this example classification can be easily performed using a linear classifier, this is not the typical case with non-synthetic sequences, with usually require a learning stage to design non-linear classifiers.

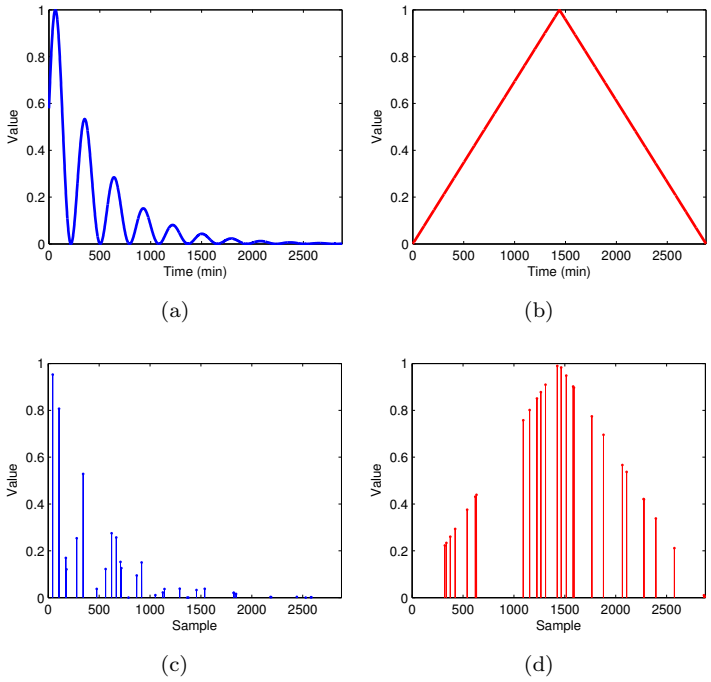


Figure 2: Time series patterns for: (a) success; (b) failure. Example of non-regularly sampled sequence from time series pattern: (c) success; (d) failure.

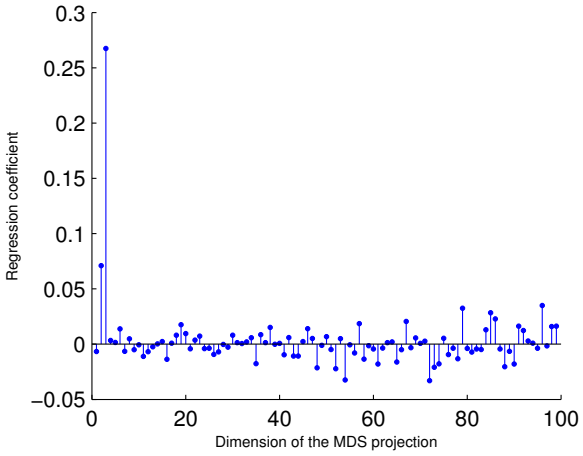


Figure 3: PLS regression coefficients after applying the PLS technique to the result of Stage 2 for the synthetic example in Section 2.4.

Marked points in Fig. 4 denote, for each class, far and close points to the linear classification boundary of maximal margin between classes. The sequences associated to these points are shown in Fig. 5. Discriminability between classes can be explained by two factors: (1) number of samples; and (2) structure. Thus, sequences with a high number of samples (Figs. 5a and 5c -successful class-; and Figs. 5b and 5d -failed class-) provide a better definition of the pattern structure they belong to and they are farther to the opposite class in the space represented in Fig. 4. On the other hand, patterns of sequences with few samples (Figs. 5e and 5g -successful class-; and Figs. 5f and 5h -failed class-) are worse represented and are also very close to the maximal margin boundary (see Fig. 4).

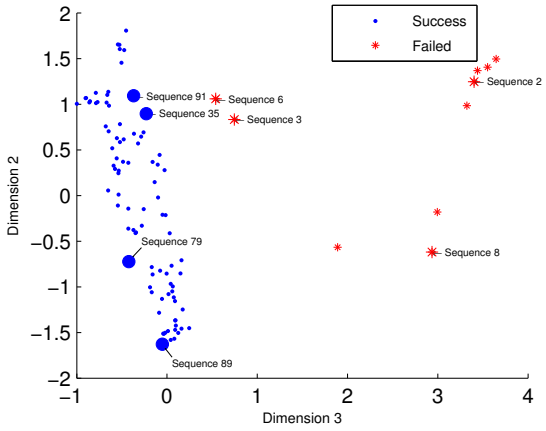
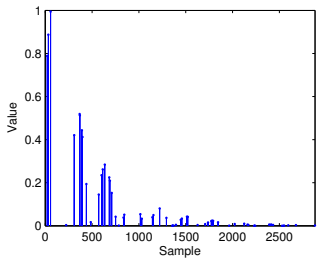


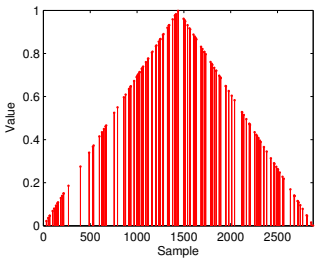
Figure 4: MDS projection of sequences in synthetic example of Section 2.4. Most influential dimensions according to the PLS algorithm have been considered. Marked points are associated to the sequences of Fig. 5.

2.5. Performance Evaluation and Validation Methods

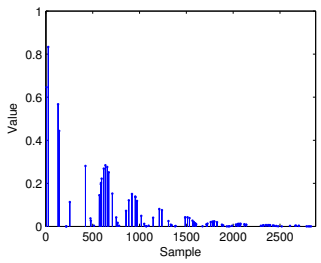
Accuracy is the most common merit figure for evaluating performance in binary classification problems. The term *baseline accuracy* is used in this paper to denote the accuracy obtained by classifying all instances as the majority class. In problems with imbalanced datasets, a deeper analysis of performance can be provided through the sensitivity and specificity. Both measures can be related



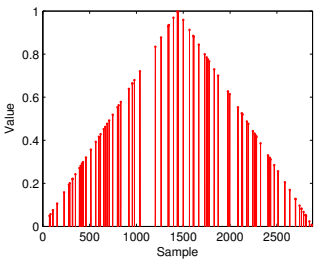
(a) Sequence 89



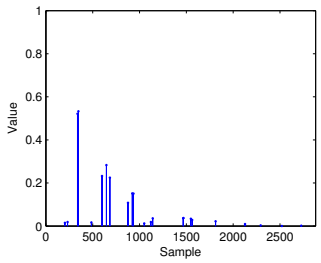
(b) Sequence 2



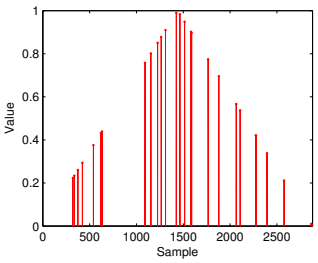
(c) Sequence 79



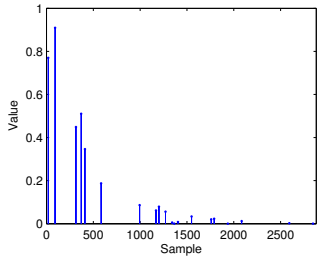
(d) Sequence 8



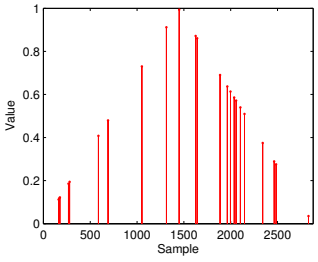
(e) Sequence 35



(f) Sequence 3



(g) Sequence 91



(h) Sequence 6

Figure 5: Time sequences associated to the points marked in Fig. 4. Successful class: left column; Failed class: right column.

through the Area Under the Curve (AUC) obtained by plotting the complementary to specificity vs sensitivity. The Balanced Error Rate (BER) merit figure is the mean of the error rate of each class and is used with imbalanced dataset because it penalizes the classification of all instances as the majority class. A lower BER indicates that both sensitivity and specificity are high and therefore performance is good.

In this work, two methods based on resampling are used for estimating the performance on an unseen test dataset:

- *Leave One Out - Cross Validation* (LOOCV) (Duda et al., 2001): this technique divides 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 (instance) 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 performance of statistical models, and therefore provides a sensible criterion for model selection and comparison.

- *Bootstrap Resampling* (Efron and Tibshirani, 1993): 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 w instances with replacement from $\mathbf{X}_w$. This resampling procedure is repeated B times to form B sets $\mathbf{X}_w^{(b)}$ $b = 1, \ldots, B$ of w instances. Bootstrap resamples $\mathbf{X}_w^{(b)}$ are conditionally independent given $\mathbf{X}_w$ and follow the same empirical distribution as x. Let us assume now that we estimate an statistic 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 properties of $\hat{\theta}_w$ can be assessed using statistics (such as standard deviation or confidence interval) of the bootstrap estimations. In this work, bootstrap resampling is used to select the

PLS classification threshold and assess the empirical distribution of the obtained merit figures.

3. Experiments and Results

3.1. Weaning Data

We selected 93 out of 253 patients under mechanical ventilation in the ICU of *Hospital Universitario Fundación Alcorcón* (Spain) from January 2010 to December 2011. 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 48 hours. A total of 104 weanings from 93 patients were considered, which were classified by the Head of the ICU into two classes: 88 for the successful weaning class (SW) and 16 for the failed weaning class (FW), yielding a value of 84.6% for the baseline accuracy. In this work, a weaning outcome corresponds to the FW class when the patient is reintubated within 48 hours after extubation (Tobin, 2006).

The weaning dataset was collected by using the clinical information system IntelliVue Clinical Information Portfolio (ICIP) by Philips. For each weaning, 18 time series (TS), 15 clinical laboratory parameters (CLP) and 12 general descriptors (GD) were collected at least once during 48 hours previous to the extubation (see Table 1). A description of these variables can be found in (Ferreira et al., 2001; Tobin, 2006; Woodrow, 2012). From a clinical point of view, these variables are potentially influential in the weaning outcome.

TS variables are characterized by a non-regular sampling (maximum sampling frequency of one sample per minute), providing a number of values per variable fluctuating between 1 and 138. Fig. 6 shows four instances of two TS variables for SW and FW classes during 48 hours before the extubation. Note that, in contrast to the synthetic example of Section 2.4, it is not evident to devise a characteristic pattern for each class.

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

Table 1: Time Series (TS), Clinical Laboratory Parameters (CLP) and General Descriptors (GD) for each weaning.

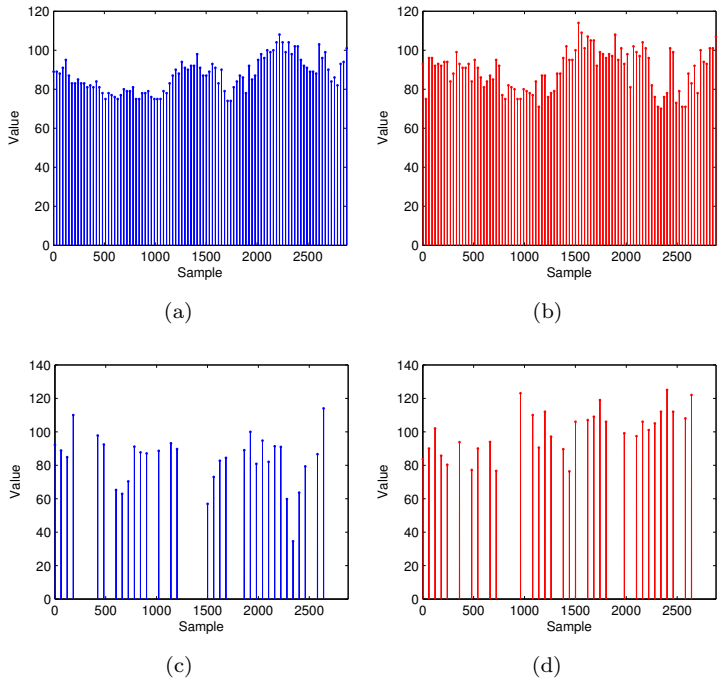


Figure 6: Sequences for two TS variables: heart rate (upper row) and expired minute volume (lower row). Left column correspond to the SW class, and right one to the FW class.

In the CLP group of variables, each variable had a reduced number of measures (up to seven, depending on the weaning), and the mean value was computed as the representative value; variables with no values (missing data) were imputed to zero. Regarding GD variables, they just have one value per variable and it can be numerical (e.g. APACHE3 index) or categorical (e.g. sex).

3.2. Conventional Tools

Three rounds of experiments were performed, two of which included some schemes proposed in other studies, such as (Giraldo et al., 2006; Hao-Yung et al., 2008; Arizmendi et al., 2009).

The first round of experiments considered features of the three kinds of variables (TS, CLP and GD). Eight statistics were obtained from each TS variable: minimum, maximum, standard deviation, variance, interquartile range, mean,

median, and summatory. A total of 187 features (8 statistics $\times$ 18 TS + 15 CLP + 12 GD) were considered. A LOOCV strategy was applied to evaluate the performance of three classifiers: linear SVM, nonlinear SVM with RBF kernel (SVM-RBF), namely $k(\mathbf{x}, \mathbf{y}) = e^{\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$, and k-NN (Duda et al., 2001; Schölkopf and Smola, 2001). Parameters of each classifier were tuned using a LOOCV strategy as well: a test sample is never used neither to design the classifier nor to tune its parameters. The search strategy was a grid-search with the SVM regularization trade off parameter $C \in [1, 150]$, the RBF-SVM kernel width parameter $\sigma \in [0.01, 15]$ and the number k of nearest neighbors used to decide the class in k-NN $k \in [1, 3, \ldots, 11]$. Baseline accuracy was not exceeded in any case. Since the number of features was larger than the number of instances, we considered two feature selection procedures: a filter method based on Mann-Whitney test (Dickinson-Gibbons and Chakraborti, 1985), and a wrapper method proposed in (Alonso-Atienza et al., 2012). In the filter method, for each feature a Mann-Whitney test was applied to test whether data supported that medians of each class were different. Those features with a p-value lower than 0.1 (following Hao-Yung et al. (2008)) were selected. Even though a false detection rate procedure (Benjamini and Hochberg, 1995) should have been applied, as we are performing 187 tests, we omitted this step as our aim was to rank features for their discrimination ability (perhaps allowing some false rejections) instead of knowing whether a feature is significant or not. The second feature selection procedure is a variant of a bootstrapped backward search using a SVM-RBF classifier. Baseline accuracy was not exceeded using selected feature sets with the three classifiers.

In the second round of experiments, each set of variables (TS, CLP and GD) were considered for predicting the weaning outcome. We included the TS trend to the previous TS statistics (i.e., the total number of features for TS was $9 \times 18 = 162$). The procedure used in the first round of experiments was applied to evaluate the performance of the same three classifiers with each set of variables. Baseline accuracy was only exceeded when TS variables were considered, yielding an accuracy of 85.57% and a BER of 47% with a linear SVM

with $C = 11$, what represents an improvement of one instance over baseline accuracy.

In the third round of experiments, other selection and classification tools available in the WEKA software (Hall et al., 2009) were evaluated with each set of variables (TS, CLP and GD). To select the most relevant features, filter (such as selection based on correlation or information gain), wrapper (with evaluators such as SVM or Multi-Layer Perceptron (MLP) and search methods such as ranker, best first or genetic algorithm) and embedded (such as C4.5 algorithm) feature selection approaches were applied using LOOCV. Selected features sets were used for weaning outcome prediction with classifiers such as decision trees (Simple CART, Random Forests and Decision Stump) or MLP, and the Adaboost M1 method with all the aforementioned base classifiers. Regarding experiments with TS variables, an accuracy[2] of 88.5% and a BER of 29% was obtained. It is interesting to remark that just four features were selected, namely interquartile range of variables heart rate, compliance and systolic blood pressure; and mean of the compliance variable, with a filter selection method based on correlation, and classified subsequently with Adaboost M1 using Decision Stump as base classifier.

From previous experiments it is clear that predicting the weaning outcome from heterogeneous data is not a simple task. Furthermore, in the above experiments, temporal reference in TS variables has not been taken into consideration. In order to deal with the raw data while maintaining the temporal reference, an investigation was made with heterogeneous TS variables, leading to the procedure proposed in Section 2.

3.3. Proposed Procedure

The NCD technique described in Section 2.1 was applied to each one of the TS variables indicated in Section 3.1. We chose the commonly used *gzip* as

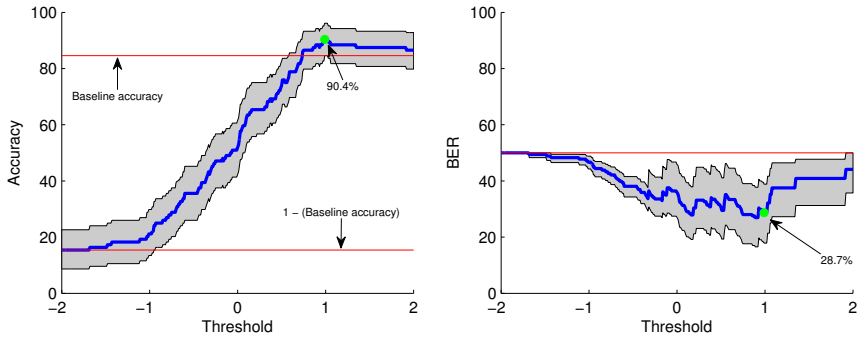
[2]This result was obtained using LOOCV for selecting features and LOOCV for evaluating performance.

compressor. A total of 18 NCD matrices of size 104×104 (number of weanings $\times$ number of weanings) were obtained, one for each TS variable. Then, the MDS technique was applied to each NCD matrix. As a result, 18 matrices of size $104 \times N_v$ were obtained, with N_v (the rank of the NCD matrix) potentially different for each TS variable. In our experiments, $N_v \in [47, 98]$.

Three classifiers (linear SVM, nonlinear SVM-RBF, and k-NN) were designed for classifying weaning outcomes considering just one TS variable. The N_v 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 would work better, PLS technique with LOOCV was applied. As indicated in Section 2.3, a procedure for thresholding the PLS prediction is necessary to perform the classification task, which we detail below. A wide enough range[3] of candidate equispaced thresholds (precision of 0.01) was considered. In order to evaluate the performance of each threshold, we bootstrapped the PLS predictions ($B = 2000$) and computed the accuracy and BER bootstrap empirical distributions, their medians and 95% confidence intervals (CIs). Threshold with the highest median accuracy (not necessarily corresponding to minimal BER) is chosen for classification. Fig. 7 shows the result of applying this procedure to the diastolic blood pressure TS variable. Note that both accuracy and BER change with the PLS threshold, reaching its maximum and minimum value, respectively, for a threshold of 0.99.

Table 2 presents the merit figures obtained for each TS variable. First and second column show the median and 95% CI of the bootstrap performance for accuracy and BER respectively. Last column in Table 2 shows the AUC, a global merit figure directly obtained from the PLS predictions of each TS variable. Best performance (boldface in Table 2) was obtained with the diastolic blood pressure variable: median accuracy of 90.4% (lower limit of the 95% CI is the baseline accuracy) and BER of 28.7% (upper limit of the 95% CI is lower than 50%). Note that this is the only case where the lower limit of the CI is at least

[3] According to the range of the corresponding PLS predictions.



(a) (b)

Figure 7: Diastolic blood pressure TS variable. Accuracy (a) and BER (b) obtained with different thresholds on the PLS prediction. Gray area represents the bootstrapped 95% CI for each merit figure, with interior line corresponding to the median. Best accuracy is obtained for a threshold of 0.99 (marked points).

equal to the baseline accuracy. The confusion matrix associated to the best performance is shown in Table 3: since the dataset is imbalanced, the SW class has a stronger influence on the PLS predictions, making it difficult classification of the FW weanings (9 false positives and 1 false negative).

Fig. 8 represents the PLS regression coefficients for the MDS features of the diastolic blood pressure TS variable. In contrast to the results of the synthetic example of Section 2.4, where one dimension stood out from the rest, now there is not clear cut to decide which dimensions are the most influential in the PLS estimation, what makes difficult an interpretation. Just for visualization purposes, Fig. 9 represents as points on a plane the instances using the two MDS dimensions with the highest influence on the prediction outcome. Absolute value of PLS regression coefficients is considered for measuring this influence, therefore dimensions 71 and 82 were selected. Note the difficulty to discriminate sequences of each class in this space. Points corresponding to sequences far and close to the cloud center (corresponding to the origin of coordinates) have been marked: sequences 23 and 37 (SW class) and sequences 50 and 70 (FW class) are more separated from the cloud center and it is expected that they have different patterns; however, as it is shown from Figs. 10a and 10c -SW class- and Figs. 10b

TS variable	**Accuracy %**	**BER %**	**AUC %**
Diastolic blood pressure	**90.4 [84.6,95.2]**	**28.9 [15.5,40.6]**	**78.9**
fiO_2	86.5 [79.8,92.3]	44.1 [34.4,50.0]	55.5
Peak airway pressure	86.5 [79.8,92.3]	44.1 [34.6,50.0]	61.5
Systolic blood pressure	86.5 [79.8,92.3]	44.4 [34.2,50.0]	52.3
Inspiratory flow	85.6 [78.8,92.3]	47.1 [40.0,50.0]	53.2
Peep	85.6 [78.8,92.3]	47.1 [39.7,50.0]	53.0
Resistance	85.6 [78.8,92.3]	47.1 [40.0,50.0]	63.6
spO_2	85.6 [78.8,92.3]	42.2 [31.2,51.1]	51.9
Inspiratory time	84.6 [77.9,91.3]	47.6 [39.5,51.6]	56.3
Compliance	84.6 [77.9,91.3]	50.0 [50.0,50.0]	56.3
Expired minute volume	84.6 [77.9,91.3]	50.0 [50.0,50.0]	51.1
Heart rate	84.6 [77.9,91.3]	50.0 [50.0,50.0]	54.3
Mean airway pressure	84.6 [77.9,91.3]	50.0 [50.0,50.0]	53.8
Plateau airway pressure	84.6 [77.9,91.3]	50.0 [50.0,50.0]	59.0
Respiratory rate	84.6 [77.9,91.3]	50.0 [50.0,50.0]	66.8
Support airway pressure	84.6 [77.9,91.3]	50.0 [50.0,50.0]	63.9
Temperature	84.6 [77.9,91.3]	50.0 [50.0,50.0]	52.8
Tidal volume	84.6 [77.9,91.3]	50.0 [50.0,50.0]	63.1

Table 2: LOOCV results for the proposed procedure when applied to each TS variable. First and second column correspond to the median and 95% CI of the bootstrapped accuracy and BER merit figures. Third column shows the AUC merit figure.

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

Table 3: Confusion matrix for the diastolic blood pressure TS variable.

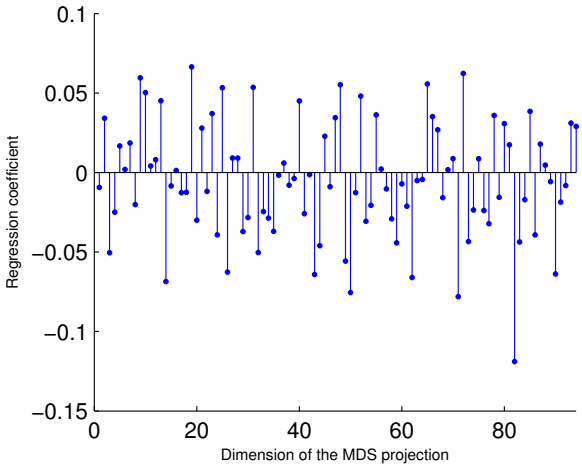


Figure 8: Regression coefficients provided by the PLS technique when applied to the MDS projections of diastolic blood pressure TS variable.

and 10d -FW class-, it is not straightforward to find the dissimilarity between patterns of different classes. On the other hand, sequences 67 and 71 (SW class) and sequences 11 and 15 (FW class) are in the middle of the cloud and it is expected that they have similar patterns; however, as it is shown in Figs. 10e and 10g -SW class- and in Figs. 10f and 10h -FW class-, it is difficult to assign similar patterns to each class. Though classification of these sequences is not evident, the proposed framework is able to detect similarities not easily captured by the naked eye neither in two nor in three dimensions, providing a reasonable solution.

4. Conclusions

A number of variables for weaning outcome prediction have been analyzed using schemes proposed in other studies. Experiments with real-world weaning data were performed using several feature selection techniques and classifiers such as decision trees, k-NN, MLP, SVM and Adaboost.

Since results with previous experiments scarcely improved the baseline accuracy, a general procedure to deal with heterogeneous time series regardless of

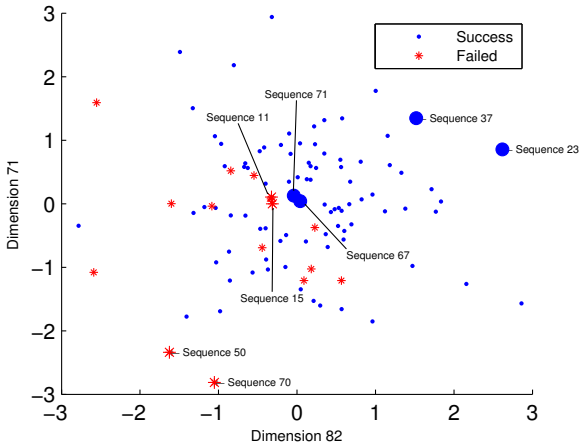
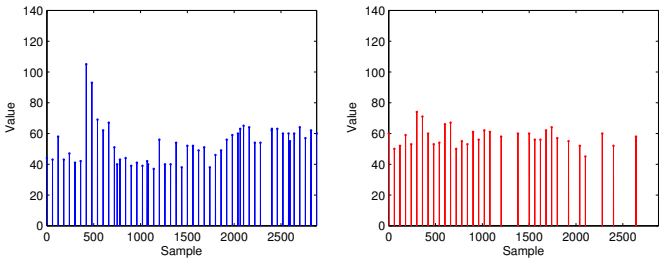


Figure 9: MDS projection of diastolic blood pressure sequences on the two dimensions with the highest influence on the PLS prediction. Marked points are associated to sequences of Fig. 10.

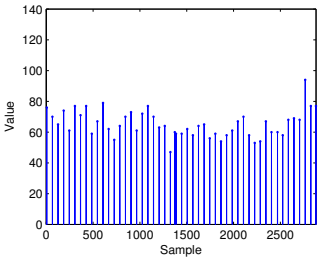
the sampling frequency and the number of samples has been proposed in this paper. The joint use of NCD and MDS allows us to provide a compact input space to design a statistical classifier. Additionally, the only parameters to be tuned are the classifier ones (in our case, the PLS threshold). Our procedure achieved the best result with the diastolic blood pressure TS variable: accuracy of 90.4% and BER of 28.7%. This represents an error rate of 9.6%, i.e. an improvement of 37% if it is compared to the physician error rate (100 - baseline accuracy = 15.4%) who classified all weanings as successful ones.

Even though previous result was achieved by analyzing TS variables one by one, its extension for considering simultaneously several TS variables is straightforward. This work has been mainly focused on time series variables, however we conjecture that performance might be enhanced by feeding the classifier with other type of variables providing complementary information for weaning outcome prediction.

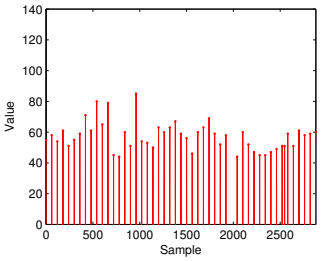


(a) Sequence 23

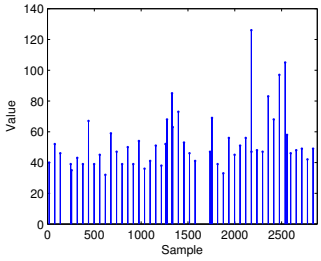
(b) Sequence 50



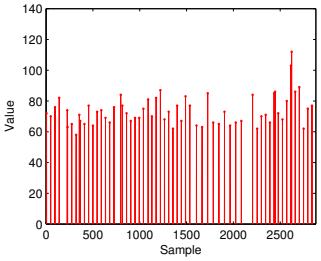
(c) Sequence 37



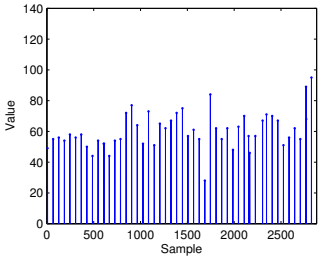
(d) Sequence 70



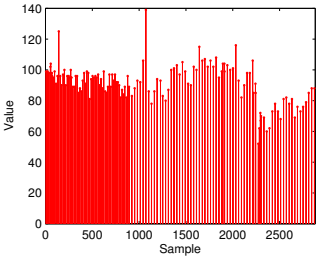
(e) Sequence 67



(f) Sequence 11



(g) Sequence 71



(h) Sequence 15

Figure 10: Associated diastolic blood pressure sequences to the marked points of Fig. 9. Successful class: left column; Failed class: right column.

5. Acknowledgements

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.

This work has been partly supported by the Spanish Government under the Research Project TEC2010-19263. Author R. Santiago-Mozos is supported by the Juan de la Cierva Program of the Spanish Ministry of Science and Innovation (Ref: JCI-2011-11150).

References

Alonso-Atienza, F., Rojo-Álvarez, J., Rosado-Muñoz, A., Vinagre, J., García-Alberola, A., Camps-Valls, G., 2012. Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. Expert Systems with Applications 39 (2), 1956–1967.

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

Axelsson, S., 2010. Using normalized compression distance for classifying file fragments. In: Availability, Reliability, and Security. pp. 641–646.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.

Bennett, C., Gács, P., Li, M., Vitányi, P., Zurek, W., 1998. Information distance. IEEE Transactions on Information Theory 40 (4), 1407–1423.

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

Bruton, A., Conway, J., Holgate, S., 1999. Weaning adults from mechanical ventilation. Physiotherapy 85 (12), 652–661.

Burns, S., Fisher, C., Tribble, S., Lewis, R., Merrel, P., Conaway, M., Bleck, T., 2012. The relationship of 26 clinical factors to weaning outcome. American Journal of Critical Care 21 (1), 52–58.

Casaseca-de-la Higuera, P., Martín-Fernández, M., Alberola-López, C., 2006. Weaning from mechanical ventilation: A retrospective analysis leading to a multimodal perspective. IEEE Trans. Biomedical Engineering 53 (7), 1330–1345.

Casaseca-de-la Higuera, P., Simmross-Wattenberg, F., Martín-Fernández, M., Alberola-López, C., 2009. A multichannel model-based methodology for extubation readiness decision of patients on weaning trials. IEEE Trans. Biomedical Engineering 56 (7), 1849–1863.

Cilibrasi, R., Vitányi, P., 2007. The google similarity distance. IEEE Trans. Knowledge and Data Endineering 19 (3), 370–383.

Dickinson-Gibbons, J., Chakraborti, S., 1985. Nonparametric Statistical Inference, 4th Edition. Marcel Dekker.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, 2nd Edition. Wiley–Interscience.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman–Hall.

Ferreira, F., Bota, D., Bross, A., Mélot, C., Vincent, J., 2001. Serial evaluation of the sofa score to predict outcome in critically ill patients. JAMA 286 (14), 1754–1758.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The weka data mining software: An update. SIGKDD Explorations 1 (1).

Hao-Yung, Y., Jiin-Chyr, H., Yung-Fu, C., Xiaoyi, J., Tainsong, C., 1–8 June 2008. 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, pp. 3981–3986.

Höskuldsson, A., 1988. Pls regression methods. Journal of Chemometrics 2, 211–228.

Jiin-Chyr, H., Yung-Fu, C., Hsuan-Hung, L., Chi-Hsiang, L., Xiaoyi, J., 2007. Construction of prediction module for successful ventilator weaning. In: Okuno, H., Ali, M. (Eds.), New Trends in Applied Artificial Intelligence. Vol. 4570 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 766–775.

Jolliffe, I., 2002. Principal Component Analysis, 2nd Edition. Springer.

Jong, S., 1998. Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18, 251–263.

Li, M., Chen, X., Li, X., Ma, B., Vitányi, P., 2004. The similarity metric. IEEE Trans. Information Theory 50 (12), 3250–3264.

Li, M., Vitányi, P., 2008. An introduction to Kolmogorov complexity and its applications, 3rd Edition. Springer.

MacIntyre, N., 2001. Evidence-based guidelines for weaning and discontinuing ventilatory support. Chest 120 (6), 375–395.

MacIntyre, N., 2004. Evidence-based ventilator and discontinuation. Respiratory Care 49 (7), 830–836.

Pinho, A., Ferreira, P., 2011. Image similarity using the normalized compression distance based on finite context models. In: International Conference on Image Processing, 18th Edition. pp. 1993–1996.

Preciado, J., Giraldo, B., 2011. Análisis y clasificación del patrón respiratorio de pacientes en proceso de retirada del ventilador mecánico. Revista Ingeniería Biomédica 5 (9), 43–49.

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

Scheinhorn, D., Chao, D., Stearn-Hassenpflug, M., Wallace, W., 2001. Outcomes in post-icu mechanical ventilation a therapist-implemented weaning protocol. Chest 119 (1), 236–242.

Schölkopf, B., Smola, A., 2001. Learning with Kernels. MIT Press, Cambridge, MA, USA.

Tobin, M., 2006. Principles and Practice of Mechanical Ventilation, 2nd Edition. McGraw–Hill.

Wold, H., 1966. Nonlinear estimation by iterative least squares procedures. In: David, F. (Ed.), Researh Papers in Statistics. Wiley, New York, pp. 411–444.

Woodrow, P., 2012. Intensive Care Nursing : A Framework for Practice, 3rd Edition. Routledge.

Yung-Fu, C., Yung-Fa, H., Jiang, X., Yuan-Nian, H., Hsuan-Hung, L., 2009. Design of clinical support systems using integrated genetic algorithm and support vector machine. In: Jiang, X., Petkov, N. (Eds.), Computer Analysis of Images and Patterns. Vol. 5702. Springer Berlin / Heidelberg, pp. 791–798.