Profiling the Amyotrophic Lateral Sclerosis disease progression through data mining techniques

Orlando Bastos Vaz
orlando.vaz@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2018

Abstract

The process of extracting information from data, which can consist of a bunch of numbers and words, is one of the most explored fields which was boosted by the ease of data storage. The use of unsupervised methods in the extraction process is a more complex task, since it is not possible to use the label of the instances to build a model for further analysis of new instances. Therefore, unsupervised clustering algorithms will be applied on a dataset of Amyotrophic Lateral Sclerosis (ALS) patients in order to analyse the need for non-invasive ventilation (NIV) of each patient. In order to do that, three different approaches were used in which the time constraint was introduced progressively. In the first approach, time was completely despised, in the second approach, the time series of every patient were summarized and, in the third one, the whole time series were used to analyse the need of NIV.

Keywords: data mining, unsupervised methods, clustering, time series, ALS.

1. Introduction

In the last few decades, the ability to learn has become an increasingly non-exclusive skill of living beings. The development of computer science led to the creation of a field known as Machine Learning (ML) whose algorithms outperformed the existing ones. Therefore, their ability to make predictions on and learn from data allowed its application in a wide a range of computing tasks.

In the beginning of the previous transformation, time-invariant data were the main focus of data collecting processes, however, the need to improve continuous processes led to the creation of new tools for gathering temporal data. As a consequence, the temporal reference within data allowed the discovery of new subfields of ML and its inherent applicability permitted the proliferation in a large number of knowledge fields. For instances, in medicine, ML has played a key role in forecasting and diagnosing domains by analysing the behaviour of human systems. Google, IBM and MIT are three institutions which have invested in the advantages of ML in medicine [1, 9, 14].

Nowadays, supervised machine learning dominates the way how data is studied and processed, however, unsupervised techniques are the most challenging ones despite their additional complexity. These methods are mainly applied when data is not labelled, however, they are still very useful, since they provide an easy way to understand how data is organized. Furthermore, they also allow the user to find new ways of splitting labelled data beyond the one already established by the class label. Despite their complexity, that currently makes infeasible their employment in a greater number of fields, unsupervised methods provide new powerful ways to search for information [21].

Based on the vast number of applications of unsupervised methods and on their relevance, some methods were applied to time series. Time series consist in a symbioses between past and recent samples which allow the detection of different behaviours in the same dataset. In result, there was an intensive expansion of time series in a large number of fields in which, for instances, past samples were used to predict future behaviours.

Therefore, time series and clustering methods will be related due to the high versatility of time series and the exalted ability of clustering to extract new information from data. Hence, after some tests, it was used a case study based on the ALS dataset which consists in multivariate time series of a set of patients. ALS in an uncommon group of neurological diseases that affects nerve cells responsible for controlling voluntary muscle movement like walking or breathing. Currently, this degenerative disease does not have a cure or a treatment to halt or reverse its progression and, thus, people usually die from respiratory failure few years after the first symptoms. During this period, the first symptoms, like speech problems (bulbar onset) or loss of strength in the arms or legs (limb onset), become more intense and the atrophy ends up spreading throughout the body [5].

This paper is structured as follows. Chapter 2 constitutes the background on application, processes and conventional algorithms. In Chapter 3, the time is taken
into account and, thus, some summarization methods and dissimilarity measures will be presented. Chapter 4 describes the case study (ALS dataset), the preprocessing steps and provides an experimental evaluation of the processes described in the theoretical chapters. Finally, the global conclusions are presented in Chapter 5.

2. Conventional Clustering Approaches

2.1. Feature Selection

Feature selection is a data mining process whose relevance and utilization has been sharply increasing in the last few years due to the massification of the stored data. The rise in the number of collected samples and the increase in the number of features per sample has hampered the implementation of these mechanisms. The main goal of feature selection methods is to elect the best subset of variables from the initial one, that can efficiently describe the input data, eliminating irrelevant variables while keeping the core information. Hence, it is possible to reduce the computing time and improve the performance [6, 20].

Filter methods, also known by Ranking ones, are one of the most desirable group of feature selection processes due to their simplicity. The inherent low computational cost and the non-use of learning algorithms (the process is not biased which avoids the overfitting) to rank the features are the most relevant advantages related to these methods [20]. Therefore, filter methods assign a score (ranking) to each variable and, based on a threshold defined by the user, some variables are validated and others are neglected.

For instance, the term-variance (TV) criterion is a filter algorithm which evaluates the ranking of the features based on the sample variance. Hence, for each variable \( X_k \), TV is defined as

\[
TV_k = \text{var}(X_k) = \frac{1}{n} \sum_{i=1}^{n} (X_{ki} - \overline{X}_k)^2,
\]

where \( \overline{X}_k \) is the mean of feature \( X_k \) and \( X_{ki} \) is the \( i \)-th sample value of the variable \( x_k \) [10].

On the other hand, the mean absolute difference (MAD) is given by

\[
MAD_k = \sum_{i=1}^{n} |X_{ki} - \overline{X}_k|.
\]

The absolute difference between the mean and the median, despite not being a measure of dispersion, it was demonstrated that it is an adequate measure to analyse the relevance of the variables [11]. Hence, median (MM) of the variable \( X_k \) is given by

\[
MM_k = |\overline{X}_k - \text{median}(X_k)|.
\]

The Inter-quartile range corresponds to the difference between the first and the third quartile. These quartiles divide an ordered dataset into four different parts and the values for each quartile correspond to the ones that divide each part (there are three quartiles). This measure of dispersion is completely insensitive to changes in the first and in the fourth quartile which also allows to be insensitive to outliers. Hence, the IQR is defined as,

\[
IQR_{R_k} = Q_{k3} - Q_{k1},
\]

where \( Q_{k3} \) and \( Q_{k1} \) are, respectively, the third and the first quartile values.

Moreover, the relation between the arithmetic mean (AM) and the geometric mean (GM) constitutes a generic dispersion measure, instead of a dispersion around the mean (like the term-variance and the mean absolute difference). This measure is given by,

\[
R_k = AM_k - GM_k,
\]

where, \( AM_k = \frac{1}{n} \sum_{i=1}^{n} X_{ki} \) and \( GM_k = (\prod_{i=1}^{n} X_{ki})^{\frac{1}{n}} \) are respectively the arithmetic and the geometric mean of variable \( X_k \). However, whether a feature has, at least, one zero occurrence, the \( R_k = +\infty \) and the criterion become useless. Hence, it was necessary to apply some functions (exponential followed by the logarithmic functions) in order to achieve the arithmetic mean geometric mean (AMGM) criterion given by,

\[
AMGM_k = \log(\sum_{i=1}^{n} e^{X_{ki}}) - \frac{1}{n} \sum_{i=1}^{n} X_{ki}.
\]

Finally, the Correlation Criteria calculates the correlation between every pair of variables using the Pearson correlation coefficient [6]. In this case, if two variables are highly correlated, it is possible to eliminate one of them, since one can easily describe the other.

In conclusion, feature selection is a crucial preprocessing approach for high-dimensional data, avoiding the curse of dimensionality, showing that the excess of information is not always good in data mining applications. The vast number of different algorithms allow the selection of features in almost every situation, improving the accuracy, for instance, of the clustering procedures. These processes will be described in the next section.

2.2. Clustering

Clustering is an unsupervised learning technique whose main goal is to find similar patterns in an unlabelled dataset. These similar patterns are discovered using a similarity measure as a criterion which analyses the distance between every observation. Subsequently, similar instances (set of elements from a dataset) are grouped into clusters resulting in a minimization of the similarity within each group and maximization of the dissimilarity between groups [18].

In Figure 1, an initial set of points is represented in the left and the result of clustering is drawn in the right side. Points with similar patterns belong to the same cluster and, thus, are identified with the same label. However, the use of this type of unsupervised techniques have some inherent problems due to the fact that information is merely extracted from data. As a result, the final
clusters sometimes are not meaningful and it is necessary to resort to another type of algorithms, not only the ones responsible for grouping the elements, but also the ones that measure the similarity between points. Despite the previous drawbacks, this technique is widely used because it allows an automatic creation of clusters, based on the interrelationships between observations from large datasets [15].

In conclusion, clustering evolves similarity measures, to inspect the closeness between observations, clustering algorithms, to group similar observations, and evaluation processes, to evaluate the quality of the separation. Therefore, these three concepts will be explained in the following sections and, for each step, some examples of algorithms will be reported.

2.2.1 Distance, similarity and dissimilarity measure

The comparison of each pair of coordinates of two observations require the use of specific methods to analyse how similar or dissimilar the observations are. In this section, examples of distance measures for continuous variables will be firstly presented and, afterwards, other examples of similarity/dissimilarity measures for categorical and mixed features will be described. Hence, the $L_p$-norms, also denominated by Minkowski distances, are one of the most well-known examples of continuous distance measures. The general formula of the $L_p$-norms is given by,

$$L_p = \left( \sum_{k=1}^{d} |x_k - y_k|^p \right)^{1/p},$$

in which $x_k$ and $y_k$ are the values of feature $k$ of observations $x$ and $y$ (multidimensional arrays), respectively, and $d$ is the dimensionality of the feature space. Although $p$ can take the value of infinite numbers, there are some particular measures that are frequently used: $p = 1$, named Manhattan or City-block distance, $p = 2$, called Euclidean distance, $p = \infty$, named Maximum or Chebychev distance [15].

Mahalanobis distance is a data-driven measure due to the fact that it depends on the dataset to which two data points belong. This dependency is related to the covariance matrix that is used to calculate the distance. The formula of this distance is given by

$$d_{Mahalanobis} = \sqrt{(x - y)^T S^{-1} (x - y)},$$

where $S$ represents the covariance matrix and $x$ and $y$ are the vectors of two observations [27].

Pearson correlation is another example of a distance measure largely used in clustering gene expression data where it tries to calculate the correlation between the shapes of two gene expression patterns. This distance is based on the Pearson correlation coefficient which is defined by,

$$r_{xy} = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2 \sum_{k=1}^{n} (y_k - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the means of $x$ and $y$, respectively, and $x_k$ and $y_k$ correspond to the variable $k$ of the observations $x$ and $y$ [25].

The cosine distance measures the angle between two non-zero vectors. This measure is totally independent of the magnitude of the vectors, since it only takes into account the angle between them. The cosine between the vectors is calculated based on the Euclidean dot product formula,

$$\cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \sqrt{\sum_{k=1}^{n} y_k^2}}$$

where $x$ and $y$ correspond to multidimensional arrays. Finally, the Spearman distance is similar to the Pearson correlation but, it assesses how well the relationship between two variables can be described using a monotonically function. Instead of analysing the relationship between the original values, it calculates the correlation between the ranks of the observations,

$$d_{Spearman} = 1 - \frac{ab'}{\sqrt{aa'} \sqrt{bb'}}$$

where $a = r_x - \bar{r}_x$ and $b = r_y - \bar{r}_y$, $r_x$ and $r_y$ are rank vectors of the features of the observations $x$ and $y$, respectively, and $\bar{r}_x = \bar{r}_y = \frac{n+1}{2}$ correspond to the mean rank. The $r_x$ and $r_y$ are the result of the replication of $r_{x_k}$ and $r_{y_k}$, $k$ times.

The Hamming distance is one of the most simplest similarity measures for categorical data. This measure calculates the percentage of different elements between observations and can be calculated using

$$d_{Hamming} = \frac{\#(x_k \neq y_k)}{n},$$

in which, $\#$ represents the number of different coordinates of $x$ and $y$.

Currently, a huge number of datasets have features of multiple types, but, using the previous metrics, it is not possible to do a joint analysis of categorical and numeric variables. The Gower distance is an example of a
mixed metric that normalizes the similarities in order to have the same contribution from every variable (weights can be used to control the relevance of some features). Therefore, the global similarity between two observations is achieved through

\[ S_{xy} = \frac{\sum_{k=1}^{n} w_{xy} S_{k}}{\sum_{k=1}^{n} w_{xy}}, \]

where \( S_{xy} \) corresponds to the contribution provided by variable \( k \) (the way how this component is calculated differs between types of variables) and \( w_{xy} \) corresponds to the weight of the previous contribution.

The distance between ordinal and continuous variables corresponds to the absolute value of the Euclidean distance normalized by the range of each variable, although the similarity corresponds to one minus the previous value,

\[ S_k = 1 - \frac{|x_k - y_k|}{r_k}, \]

where \( r_k \) represents the range of the variable \( k \). Moreover, the similarity between two nominal variables is a dichotomous process, since it gives the value of 1 when \( x = y \) and the value of 0 when \( x \neq y \), but the similarity between binary variables is calculated differently. The main difference between the nominal and the binary approaches is the absence of value in the binary approach \( (x_k = 0 \text{ and } y_k = 0) \) which is not taken into account \( (w_{xy} = 0) \).

Despite of all, the choice of the best similarity measure is a difficult task because the performance of each metric may depend on the type of the data, the goal of the application, the field of the data or even the dimensionality of the one. Many researchers have been attempting to find the most appropriate distance measure for each dataset but, it still is a big challenge. Hence, it is usually necessary to try more than one metric in order to conclude about the best option for that context [25].

### 2.2.2 Clustering Algorithms

The distance measures have an important role in comparing every instance in order to evaluate the similarity/dissimilarity between them. The next step consists in grouping the observations in different clusters based on the information gathered by the distance measures. This is performed by a clustering algorithm. The main goal of these algorithms is to arrange the multiple instances in two or more clusters and, in the same cluster, the observations must be similar as much as possible and, on the contrary, in different clusters, the observations must be as different as possible [26].

Although this common goal characterizes all the clustering algorithms, the way how they deal with the observations varies between them which often leads to different final results. Hence, the partitioning and hierarchical categories of clustering methods will be described. Beyond these two examples, density, grid and model-based categories are the other types of clustering methods that are not going to be explored, however some information about them can be seen in [2, 4, 18, 26].

Partitioning methods are the ones where the \( n \) objects of the dataset are split into \( k \) clusters (parameter chosen by the user), but, every cluster must have, at least, one element \( (k \leq n) \). Particularly, these methods perform an iterative optimization that reallocates the objects to the clusters in order to minimize the objective function [4]. The k-means and k-medoids are two examples of partitioning methods.

The hierarchical clustering methods create trees of clusters in order to make a connection between every observations [18]. This kind of methods can be divided into two groups: agglomerative and divisive methods. Agglomerative are the ones that considers that each object is a distinct cluster and, then, the algorithm merges the clusters, creating new and bigger ones. This process finishes either when there is a single cluster or when a stopping criterion is reached (for instance, this criterion can be the desired number of clusters). In this methods, besides the distance measures, a linkage criteria is also used as a parameter, since they regulate how the metric should be used to calculate the similarity between clusters. Single, complete, average and ward linkage are four examples of linkage criteria.

On the other hand, the divisive algorithms start with only one cluster which is split into smaller ones. This algorithm stops when each observation corresponds to a single cluster or when the stopping criterion is reached. DIvisive ANAlysis clustering (DIANA) is an example of this type of algorithms[24].

### 2.2.3 Clusters Validity

The panoply of clustering algorithms allows the use of these techniques in multiple fields and therefore, numerous clusters are created with different characteristics. However, some sets of clusters are not meaningful which can be caused by a non-adequate use of the clustering algorithms or by the non-structured datasets that are used. Hence, it is mandatory to evaluate the quality of the clusters using a suitable method, since the visualization of a multidimensional domain is a difficult task for humans [13]. These methods must be objective and should not depend on the applied clustering algorithm, however, they should provide information about the quality of the clusters. Therefore, there are two different approaches that can be performed in the cluster evaluation step: external and internal [2].

The external index compares the clustering results to the ground truth and, thus, these procedures are categorized as supervised processes. This means that the results are compared to a pre-specified structure that has information about the class labels. Accuracy, sensitivity, specificity and balanced accuracy are four examples of external indexes which are calculated based on the true
3. Temporal Clustering Approaches

3.1. Summarization Methods

The time series, TS, summarization methods are processes able to describe a TS using its global characteristics. These methods are applied over an univariate TS with multiple time points and, thus, the application of each method results in a single value. However, this transformation usually evolves the loss of relevant information. Hence, some summarization methods are usually applied simultaneously in order to transform an univariate TS into a feature vector in which every value is saved [22].

The mean corresponds to the sum of the values of the whole univariate TS divided by its length (number of time points). The mode corresponds to the most frequent value in the TS and the median is middle value when the elements of a TS are organized in an ascending order (if the number of elements is a multiple of 2, the median corresponds to the mean between the two middle values). The variance represents the amount of variation around the mean value and it is defined as,

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{L-1}. \quad (15)$$

The fractal dimension-based summarization is the type of method that provides an index of complexity by comparing the changes in the details to the variations in scale. In this case, the fractal dimension is going to be applied on TS (1-D). The variance will be used as self-similarity function which is given by,

$$Var = \frac{\sum_{d=1}^{N(d)} (x(t_i) - x(t_{i+d}))^2}{2N(d)}, \quad (16)$$

where $x(t_i)$ is the value of the TS at time $t_i$ and $t_{i+d}$ is other value of the TS with distance $d$ from instance $t_i$. Moreover, $N(d)$ is the total number of points between $t_i$ and $t_{i+d}$.

In order to calculate the fractal dimension,

$$FD = \frac{d - s}{2}, \quad (17)$$

it is necessary to calculate the slope, $s$ that results from the graph of the log of the variance versus the log of the distances (semi-variogram).

The similarity or dissimilarity between two TS is one of the most important procedures to cluster the observations from a dataset. The possibility of finding not only perfect matches between TS, but also approximate matches between them makes this phase critical for a posterior application of a clustering algorithms [2, 17].

In Section 2.2.1, some distance measures were enumerated to detect differences between simple observations that do not have time indexaion. Furthermore, those distances can also be applied to univariate TS. For instance, the Euclidean distance is used to calculate the distance between two TS.

One of the most acclaimed algorithms is the Dynamic Time Warping (DTW). DTW is a generalization of classical methods which can align discrete sequences of continuous values of different length in order to minimize the distance between sequences [12, 18].

The temporal correlation and raw values method is an adaptive dissimilarity measure that results from a symposium between the dissimilarity on raw val-
ues and the dissimilarity on temporal correlation behaviour [23]. The expression, $d_{CORT}(x, y) = \Phi(CORT(x, y))$ reflects the distance between TS $x$ and $y$ whose meaning can be divided into three different parts. The $CORT(x, y)$ measures the proximity between the dynamic behaviours, which means that it analyses whether TS have similar growths in direction and rate at the same instant of time using the following expression:

$$CORT(x, y) = \frac{\sum_{t=1}^{T-1} (x_{t+1} - x_t)(y_{t+1} - y_t)}{\sqrt{\sum_{t=1}^{T-1} (x_{t+1} - x_t)^2 \sum_{t=1}^{T-1} (y_{t+1} - y_t)^2}},$$

(22)

where $T$ is the length of the TS and $x_t$ and $y_t$ are the values of the TS in the instant $t$. The value of $CORT(x, y)$ belongs to $[-1, 1]$, where 1 represents similar dynamic behaviour (similar growth in direction and rate) and −1 indicates similar growth in rate but in opposite directions. Moreover, $\Phi[u] = \frac{1}{1 + e^{-u}}$ is an adaptive tuning function whose parameter $k$ controls the weight of the dissimilarity between dynamic behaviours (if $k = 0$, $d(x, y) = \delta(x, y)$). Finally, $\delta(x, y)$ is the function that measures the distance between the raw values of both TS. This function can be, for instance, the Euclidean or the DTW distance [7].

On the other hand, some correlation-based distances are directly related to the Pearson’s correlation criteria which is given by

$$COR(x, y) = \frac{\sum_{t=1}^{n} (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{n} (x_t - \overline{x})^2 \sum_{t=1}^{n} (y_t - \overline{y})^2}},$$

(23)

where $\overline{x}$ and $\overline{y}$ are the average values of the serial realizations $x$ and $y$. Hence, using the previous criteria, the following metrics constitute two examples of cross-correlation-based distances:

$$d_{CORRI}(x, y) = \sqrt{2(1 - COR(x, y))},$$

(24)

and

$$d_{CORR2}(x, y) = \sqrt{\left(\frac{1 - COR(x, y)}{1 + COR(x, y)}\right)^2 \beta \geq 0},$$

(25)

where $\beta$ is the parameter that allows the regulation of the fast decreasing of the distance between two TS [23]. Equation (24) is used when $\beta$ is not defined and, otherwise, Equation (26) is used.

The autocorrelation is another example of a dissimilarity measure that performs a weighted Euclidean distance between the autocorrelation coefficients of two distinct TS. These coefficients are given by $\rho_x = (\hat{\rho}_{1,x}, \hat{\rho}_{2,x}, \ldots, \hat{\rho}_{L,x})$ and $\rho_y = (\hat{\rho}_{1,y}, \hat{\rho}_{2,y}, \ldots, \hat{\rho}_{L,y})$ (TS $x$ and $y$, respectively) and $\rho_{x} \approx 0$ and $\rho_{y} \approx 0$ for $i > L$ [23]. Hence, the distance between $x$ and $y$ are given by

$$d_{ACF}(x, y) = \sqrt{(\rho_x - \rho_y)^T \Omega (\rho_x - \rho_y)},$$

(26)

where $\Omega$ corresponds to the matrix of weights.

The discrete wavelet transform (DTW) is essentially a representation method that can be used as an intermediate step to analyse the similarity between TS. First, the original TS are replaced by the respective wavelet approximation coefficients in an appropriate scale and, then, the dissimilarity is measured between those coefficients using the Euclidean distance. This method provides a way of converting TS from the time-domain into the time-frequency domain through a scale-wise decomposition, keeping most of their energy constant. Hence, a TS, $x$ with length $L$, located in a scale $J = \log_2 L$ is decomposed in a specific scale $j \in [0, 1, \ldots, J - 1]$, whose coefficients are given by $H_j(x) = \{A_j, D_j, D_{j+1}, \ldots, D_{J-1}\}$. The $A_j$ coefficients are considered the approximation ones and $D_j, \ldots, D_{J-1}$ are the wavelet coefficients (represent the detailed information of $x$ which is also associated to high-frequency). Since noise often exists in high-frequency part of the signal, only the first few coefficients ($A_j$) are used to analyse the dissimilarity. Moreover, the scale $j$ corresponds to the one that achieves the maximum dimensionality reduction and, simultaneously, that preserves the maximum amount of information from the original data [28, 23]. Hence, the DWT calculates the dissimilarity of two TS as

$$d_{DWT}(x, y) = \sqrt{\sum_{k} (a^x_{k,j} - a^y_{k,j})^2}.$$

(27)

The symbolic aggregate approximation (SAX) is a symbolic representation of a TS which first reduces the dimensionality of the data and, then, calculates the distance between the TS [19]. This distance measure outperforms the remaining symbolic dissimilarity measures, since it is able to reduce the dimensionality and has high correlation with the distance measures calculated with the original data. This dissimilarity measure starts by normalizing the TS to zero mean and unit standard deviation. The second step consists in reducing the dimensionality of the data using piecewise aggregate approximation (PAA). After the dimensionality reduction step, the vector with the mean values is transformed into a string using the Gaussian curve and the number of symbols initially defined. After the symbolic representation, the distance between the strings is calculated using the following expression:

$$dist_{R,T}(r, c, \beta_{\max(r,c)} - \beta_{\min(r,c)}, \mu, \nu) \begin{cases} 0, & \text{if } |r - c| \leq 1. \\ \beta_{\max(r,c)} - \beta_{\min(r,c)}, & \text{otherwise}. \end{cases}$$

(28)

where $r$ and $c$ are the cells from the lookup table of Figure 2 [19].

The compression-based dissimilarity measure (CDM) is a parameter free approach that indicates the amount of common patterns of two strings.

$$d_{CDM}(x, y) = \frac{C(xy)}{C(x) + C(y)}.$$

(29)
In this section, the dataset was transformed into a panel data with 1671 independent multivariate observations in order to estimate the class of each time point based only on the values of the variables that belong to it without recourse to the patient’s history. In brief, two main approaches were taken into account in order to study the influence of the dynamic and static data: in the first approach every feature was considered, but in the second one only the time-variant ones were taken into account. In both cases two different clustering algorithms (partitioning and hierarchical) were applied. In the first approach, the Gower distance and the Hamming distance (after an equal-width and equal-frequency discretization of the non-categorical features) were used to compare the observations. These dissimilarity measures were used with PAM, DIANA and the hierarchical agglomerative algorithms (Ward and Average links).

Table 1 shows the results of the application of the Gower distance with the clustering algorithms. Based on it, it is possible to see that the sensitivity and specificity values are both either low or opposite (when sensitivity is high, specificity is low or vice-versa) which results in a low balanced accuracy in every situation. Moreover, the Silhouette also presents a low value (it is not close to one) which means that the clusters demonstrate either a proximity between the clusters or a bad compactness within the clusters. In the second test the data were discretized and, then, the Hamming distance was used to cluster the data.

4.1 Time Independent Analysis

In this section, data is not going to be used as a sequence of consecutive elements with a time relationship. In the first part of the section, the time points will be considered independent from each other and, in the second one, the behaviour of the whole TS is going to be summarized into a few new variables. Hence, conventional similarity measures and clustering algorithms were used to cluster the data.

4.1.1 Conventional clustering with the imputed dataset

In this section, the dataset was transformed into a panel data with 1671 independent multivariate observations in order to estimate the class of each time point based only on the values of the variables that belong to it without recourse to the patient’s history. In brief, two main approaches were taken into account in order to study the influence of the dynamic and static data: in the first approach every feature was considered, but in the second one only the time-variant ones were taken into account. In both cases two different clustering algorithms (partitioning and hierarchical) were applied.

In the first approach, the Gower distance and the Hamming distance (after an equal-width and equal-frequency discretization of the non-categorical features) were used to compare the observations. These dissimilarity measures were used with PAM, DIANA and the hierarchical agglomerative algorithms (Ward and Average links).

Table 1 shows the results of the application of the Gower distance with the clustering algorithms. Based on it, it is possible to see that the sensitivity and specificity values are both either low or opposite (when sensitivity is high, specificity is low or vice-versa) which results in a low balanced accuracy in every situation. Moreover, the Silhouette also presents a low value (it is not close to one) which means that the clusters demonstrate either a proximity between the clusters or a bad compactness within the clusters. In the second test the data were discretized and, then, the Hamming distance was used to cluster the data.

4. Results

In this section, the previous methods will be applied on a real dataset dataset that has information on 630 patients collected at Hospital de Santa Maria in Lisbon. The information consists of a set of multivariate time series including 32 features of which some correspond to the results of the medical exams and the remaining ones constitute the description of each patient.

This dataset had a huge number of missing values and six of the features had more than 70% of missing values. Thus, some preprocessing steps were applied in order to impute them. First, horizontal imputation was used to fill some of the missing values and, after this imputation, every patient with more than 50% of missing values were eliminated. In the next step, every feature with more than 30% of missing values were also eliminated and, finally, vertical imputation was used to fill in the remaining blanks.

where \( x \) and \( y \) are two distinct strings, \( C(x) \) and \( C(y) \) are the size of the file with the compressed strings and \( C(xy) \) is the size of the file with the compressed concatenated strings [16].

Similarly, the normalized compression-based dissimilarity (NCD) is another dissimilarity measure that follows the same rules as the previous one. This metric is based on the idea that whether two files are similar, considering a certain feature that is described by a specific metric, the files are also similar when it is used a normalized information metric [23, 8]. Therefore, the expression of this metric is given by

\[
d_{\text{NCD}}(x,y) = \frac{C(xy) - \min(C(x),C(y))}{\max(C(x),C(y))},
\]

where \( ED \) corresponds to the Euclidean distance and,

\[
C_{\text{FD}}(x,y) = ED(x,y) \times CF(x,y),
\]

where \( ED \) corresponds to the Euclidean distance and,

\[
CF(x,y) = \frac{\max(CF(x),CF(y))}{\min(CF(x),CF(y))},
\]

where,

\[
CF(x) = \frac{\sum_{i=1}^{L-1} (x_i - x_{i+1})^2}{L},
\]

is the complexity estimation of a TS. where \( x_i \) and \( x_{i+1} \) correspond to consecutive elements of a TS \( x \) with length \( L \). Finally, this distance measure is applicable to TS with the same sampling rate, the same number of observations and with normalized amplitudes [3].

4. Results

In this section, the previous methods will be applied on a real dataset dataset that has information on 630 patients collected at Hospital de Santa Maria in Lisbon.
was used to calculate the dissimilarities.

<table>
<thead>
<tr>
<th></th>
<th>Sens</th>
<th>Spec</th>
<th>BAcc</th>
<th>Silh</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM</td>
<td>0.618</td>
<td>0.488</td>
<td><strong>0.553</strong></td>
<td>0.235</td>
</tr>
<tr>
<td>DIANA</td>
<td>0.502</td>
<td>0.488</td>
<td>0.508</td>
<td>0.219</td>
</tr>
<tr>
<td>Hier. Ward</td>
<td><strong>0.632</strong></td>
<td>0.484</td>
<td><strong>0.552</strong></td>
<td>0.246</td>
</tr>
<tr>
<td>Hier. Average</td>
<td>0.013</td>
<td><strong>0.999</strong></td>
<td>0.506</td>
<td><strong>0.296</strong></td>
</tr>
</tbody>
</table>

Table 1: Results of the application of PAM, DIANA and hierarchical agglomerative algorithms with Gower distance on the imputed dataset.

As was previously referred, the second test of the first approach consists in discretizing the continuous and ordinal variables, which transformed a mixed dataset into a categorical one. PAM, DIANA and the hierarchical algorithm with ward and average links were applied with the Hamming distance which resulted in high specificities and, in the opposite, low sensitivities. Thus, balanced accuracies are, in general, around 0.5 which represents a mixture between observations of different classes (Silhouette values are also low).

On the other hand, the second approach in which only time-variant features were used in three different ways: raw data, normalized data (subtracts the mean and divides the result by the standard deviation) and discretized data. In the first two tests, k-means, PAM, DIANA and the hierarchical agglomerative algorithms (ward and average links) were used to group data based on the $L_p$-norms, correlation, cosine, Mahalanobis and Spearman distances and, in the third test, the same clustering algorithms were used with the Hamming distance.

<table>
<thead>
<tr>
<th></th>
<th>Sens</th>
<th>Spec</th>
<th>BAcc</th>
<th>Silh</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>City-block</td>
<td>0.7246</td>
<td>0.5551</td>
<td><strong>0.6399</strong></td>
</tr>
<tr>
<td>DIANA</td>
<td>City-Block</td>
<td>0.4967</td>
<td>0.7484</td>
<td>0.6225</td>
</tr>
<tr>
<td>Hier. Average</td>
<td>Chebyshev</td>
<td>0.3393</td>
<td>0.8238</td>
<td>0.5815</td>
</tr>
<tr>
<td>K-means</td>
<td>City-block</td>
<td>0.6902</td>
<td>0.6277</td>
<td>0.6589</td>
</tr>
<tr>
<td>DIANA</td>
<td>City-Block</td>
<td>0.0093</td>
<td><strong>0.9934</strong></td>
<td>0.5164</td>
</tr>
<tr>
<td>Hier. Average</td>
<td>Chebyshev</td>
<td>0.0148</td>
<td>0.9934</td>
<td>0.5041</td>
</tr>
<tr>
<td>K-means</td>
<td>City-block</td>
<td>0.0131</td>
<td><strong>0.9934</strong></td>
<td>0.5033</td>
</tr>
</tbody>
</table>

Table 2: Partial results of the application of K-means, PAM, DIANA and hierarchical agglomerative algorithms with multiple distance measures on time-variant features.

In general, the hierarchical average (raw and normalized) and the DIANA (normalized) algorithms have the lowest balanced accuracy values (around 0.5) and it is possible to say that there is a huge cluster with almost every element from both classes and a small one with the remaining ones, meaning that the classes are not well separated.

In the test with raw data, specificity and sensitivity of these values are very low which results in a reduced balanced accuracy. The higher balanced accuracy (0.6399), which represents the best separation of classes, is reached using k-means with the city-block distance measure. Despite being the highest value, it is not a great value which shows that the classes are roughly separated. Hence, it is possible to conclude that the classes are still very intertwined in every situation. On the other hand, characteristics of the clusters built with normalized data are similar to the previous ones, however, the Silhouette values are usually lower. Finally, the clusters built with the discretized data demonstrate the same behaviour as the ones described in the previous two paragraphs, since the balanced accuracies are relatively low and, in addiction, the Silhouette values are very close to zero which confirms that the results are not meaningful.

### 4.1.2 Conventional clustering with summarization methods

Summarization methods are used to transform a multivariate TS into a single time-invariant multivariate instance, eliminating the concept of time. Each summarization method converts an univariate TS into a single instance and, thus, in this test, multiple methods were used in order to gather more information from the TS. Therefore, mode (every type of variable), mean, median, mode, variance (integer and non integer variables), FD, SRE, LRE, SK and KUR (integer variables) were applied and, after creating new variables, conventional clustering techniques were performed on the summarized data (partitioning, PAM, hierarchical agglomerative with Ward and divisive, DIANA).

The summarization methods were applied on a set of datasets that were built to test the performance of these methods. These new datasets use different amounts of temporal information and use the labels of different time points. For instance, these methods were applied on data from time points one and two but the class of time points two, three, four, five and six were used in the validation process (if the class of time point four is used all the patients must have a TS with a length equal or longer than four). Hence, the goals of this test consist in verifying the influence of the amount of information using the same class in the validation process and to analyse the quality of the forecast some time points ahead, using the same amount of information but different labels in the validation process.

In Figure 3, it is possible to see an example of the variation of the sensitivity, specificity and balanced accuracy with the amount of information involved and with the class after the application of the summarization methods and the hierarchical algorithm. In both figures, the main difference between the points of a single curve is the number of patients and the estimated class and the key difference between curves falls on the amount of temporal information (number of time points) used by the summarization methods. For instance, the red line starts in C2 due to the fact that it uses information...
from the first two time points and the validation process is performed with class of the second, third, fourth and sixth time points. On the other hand, the purple line starts in the fifth time point because it has information about the first five time points and the validation process is computed with the class value of the fifth and the sixth time point.

![Hierarchical Ward](image)

Figure 3: Variation of the balanced accuracy, sensitivity and specificity with the amount of information and with the class used in the validation process.

It is possible to see that there is a mixture between the solid and the dashed lines, meaning that there is an apparent similarity between the sensitivity and specificity values. The purple and the blue lines have the highest accuracy values in C5 and C6, respectively, and both attach greater importance to the sensitivity. Thus, the true positives are well grouped (sensitivity is 0.85 in both), but the number of false positives is really high due to the reduced specificity (less than 0.53), meaning that the classes are not well separated. Furthermore, when C3 and C4 are used, the green and yellow points, respectively, are the ones with the highest accuracies and their specificities are, approximately, 0.74 and 0.68, respectively. However, in both cases, the patients that need NIV are divided between the clusters due to their low sensitivity.

Based on the previous analysis, it is possible to conclude that the previous clustering algorithm is not able to do a good estimation of every class, specially C2. This conclusion can be obtained from the three tests (with PAM, DIANA and Hierarchical Ward).

4.2. Time Dependent Analysis

Since the beginning of this chapter, there has been a progressive introduction of the temporal component in the analysis of the ALS dataset. In Section 4.1.1, the temporal component was completely despised. However, in the Section 4.1.2 there was a subtle introduction of the temporal parcel due to the fact that the main focus fell on the global characteristics obtained from each TS. In this final subsection, a time dependent analysis was performed in which the temporal relationship within features were preserved. Hence, it was possible to compare, not only the absolute values of the features, but also their variations with time.

The choice of the most suitable dissimilarity measure was computed under well-known circumstances, since they were tested on synthetic datasets and a statistical analysis was performed in order to choose the best metrics. Hence, 6 datasets were created in order to analyse the changes in performance caused by modifications of the slope, the y-intercept and the linear variation. Additionally, 4 datasets were built to analyse situations in which variations fall on the period of the signal and on the offset and other 3 datasets have random signals which were generated with different probability distributions and with different offsets. The remaining datasets consist in a mixture of the previously described characteristics.

After building the datasets, some dissimilarity measures were applied on the data: Euclidean distance (Eucl), dynamic time warping method (DTW), temporal correlation and raw values method with Euclidean and DTW (CortEucl and CortDTW, respectively), correlation-based method (Corr), autocorrelation-based method (Autocorr), discrete wavelet transform method (DWT), symbolic aggregate approximation measure with alphabet size of 3 and 4 (SAX3 and SAX4, respectively), model-based dissimilarity measure proposed by Piccolo (Picol), compression-based dissimilarity method (Compr), normalized compression distance (NormCompr) and complexity-invariant distance (CompleInv). Moreover, three different clustering algorithms were used to group the observations based on the outputted dissimilarity matrices: Agglomerative Hierarchical with two different linkage criteria, Ward and Average, DIANA and PAM.

Friedman and Nemenyi tests were applied on the accuracy values, allowing not only a general analysis, but also a pairwise one. The Nemenyi test showed that some pairs of the algorithms have a p-value lower than 0.05, meaning that some algorithms are not equivalent to each other. Therefore, the distance measures considered significant were further used to cluster time series from a real dataset.

In order to perform the temporal analysis, two datasets were built based on the original one (ALS dataset): one was truncated to five and other to six time points. Hence, it was possible to analyse the influence of the length of the TS and the effects resulting from de application of different methods. In order to compare these TS, the dissimilarity measures considered statistically significant in the previous study were used. However, the ALS dataset is constituted by multivariate TS of multiple patients and the dissimilarity measures studied were only suitable for univariate TS. Hence, it was necessary to extrapolate by disintegrating the multivariate TS into multiple univariate ones which correspond to the evolution in time of a single feature. Then, the univariate TS of the same feature of every patients were compared using the dissimilarity metrics. For each fea-
ture, a dissimilarity matrix with the distances between every pair of TS was outputted and, in the end, all dissimilarity matrices were summed. Hence, this results in a single dissimilarity matrix whose values correspond to the sum of the distances between every univariate TS of each pair of patients. Then, using the clustering algorithms where the respective distance measure was considered statistically significant, the observations were merged in clusters.

In Table 3, the results show that, in the dataset with five time points, the best separation of classes was reached by the PAM algorithm using the CID (dissimilarity measure with the best results in the Nemenyi test), since it has the highest balanced accuracy (0.620). In this case, the specificity value is high (0.746), but the sensitivity one is considerably low (0.495). However, it has the second lowest Silhouette, meaning that the clusters are not either compact or separated (or both). The hierarchical average with CID metric presents the highest sensitivity (1.0) and Silhouette (0.641) values, however, the classes are very mixed (balanced accuracy is 0.509) and the specificity value is almost zero.

The results of the dataset with six time points show that the best separation of classes occurs (balanced accuracy is 0.643), coincidentally, with the same clustering algorithm (PAM) and the same dissimilarity measure (CID). In this case, the balanced accuracy value is lightly superior than in the other dataset because of the increase in the sensitivity one (0.588).

<table>
<thead>
<tr>
<th></th>
<th>Sens</th>
<th>Spec</th>
<th>BAcc</th>
<th>Silh</th>
</tr>
</thead>
<tbody>
<tr>
<td>5TP</td>
<td>0.495</td>
<td>0.746</td>
<td>0.620</td>
<td>0.067</td>
</tr>
<tr>
<td>Hier. Average</td>
<td>1.0</td>
<td>0.018</td>
<td>0.509</td>
<td>0.641</td>
</tr>
<tr>
<td>6TP</td>
<td>0.588</td>
<td>0.697</td>
<td>0.643</td>
<td>0.167</td>
</tr>
<tr>
<td>Hier. Average</td>
<td>0.985</td>
<td>0.0</td>
<td>0.493</td>
<td>0.161</td>
</tr>
</tbody>
</table>

Table 3: Results of the application of PAM and hierarchical agglomerative algorithm with CID measure on multivariate dataset.

5. Conclusions

Data mining has been a very important tool for the society and its role solving problems has been increasing a lot in the last years. Specifically, the unsupervised methods are usually the most challenging ones, since they are usually used in an exploratory way when the information about the data is reduced. Moreover, these methods are also used to explore more about the relevance of the features.

The first approach was performed in order to decide whether it was possible or not to choose if a patient needs to be treated with NIV based on the medical exams of a single time point. However, the results showed that it is not possible to make a decision due to the low silhouette and balanced accuracy values of most of the tests.

The summarization methods also do not provide good results about the separation of classes, however, there is a test whose balanced accuracy was boosted by sensitivity and specificity values (around 0.7). This means that 70% of the NIV and non-NIV patients were correctly assigned to the right cluster which is close to a good separation of classes.

Finally, the last approach did not also have good results, however, it was possible to see that the number of patients had more impact than the amount of temporal information. Despite this conclusion, the final results demonstrate that, in both datasets, the separation of classes is not good enough.

References