ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Interactive visual analytics for medical data: application to COVID-19 clinical information during the first wave

José M. Enguita[1], Diego García[1] María D. Chiara[2], Nuria Valdés[3], Ana González[1], Abel A. Cuadrado[1], and Ignacio Díaz[1] *

1- University of Oviedo - Dept of Electrical Engineering
Edificio Torres Quevedo, módulo 2, Campus de Gijón 33204 - SPAIN

2- Institute of Sanitary Research of the Principado de Asturias
Hospital Universitario Central de Asturias, Oviedo 33011 - SPAIN.

3- Department of Internal Medicine, Section of Endocrinology and Nutrition
Hospital Universitario de Cabueñes, Gijón 33204 - SPAIN.

Abstract. Biomedical data recorded as a result of clinical practice are often multi-domain –involving lab measurements, medication, patient attributes, logistic information–, and also highly unstructured, with high rates of missing data and asynchronously sampled measurements. In this scenario, we need tools capable of providing a broad picture prior to more detailed analyses. We present here a visual analytics approach that uses the morphing projections technique to combine the visualization of a t-SNE projection of clinical time series, with views of other clinical or patient's information. The proposed approach is demonstrated on an application case study of COVID-19 clinical information taken during the first wave.

1 Introduction

COVID-19 pandemics has been an unprecedented situation with serious medical, economic and social consequences. Along the crisis, large amounts of clinical data were generated and made public in an effort to acquire valuable knowledge to deal with an unbearable number of hospital admissions, with a significant number of critically ill patients. However, clinical data are in most cases highly unstructured, and this worsened as a result of the emergency of the moment. While after two years of pandemics we have improved our knowledge and tools to face COVID-19, new variants and even other kind of pandemics could emerge, and tools are still needed to enable rapid, efficient and intuitive analysis of COVID-19 behavior in different scenarios, such as response to treatments, clinical conditions and patient groups, as well as decision support and protocol development. These tools should allow for fast and interactive multilevel exploratory analysis –ranging from the "big picture" to detailed and quantitative analysis– and be powered by machine learning algorithms that can reveal hidden patterns in the data and integrate this information in a visual and intuitive way.

*This work is part of Grant PID2020-115401GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.
Demo video: `https://youtu.be/IgX5FqK3Cms`

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Recently, a highly interactive approach for biomedical data analysis called *Morphing Projections* (MP) was proposed [1], that proved to be useful to discover useful knowledge by cross-analyzing genomic and clinical data. In this paper, we propose the application of the MP approach, including time series dimensionality reduction, to carry out a comprehensive analysis of a large public dataset "Covid Data Saves Lives" with a large amount of time-referenced COVID-19 biomedical data.

2 Materials and methods

2.1 Covid data saves lives

On April 15, 2020, HM Hospitales made public a dataset with all the clinical information on patients treated in their hospital centers in Madrid for the SARS-CoV-2 virus. This dataset, available to the international scientific community under request, was given the name "Covid Data Saves Lives"[1] and contains the anonymized records of 2,310 patients, admitted with a diagnosis of covid positive or covid pending, since the beginning of the epidemic to the date it was made public (corresponding to the first wave). It contains records on diagnosis, vital signs at the time of admission, treatments, intensive care unit (ICU) admissions, reason for patient discharge, and many more, where some of the attributes were asynchronously obtained several times for each patient along the follow-up, bearing highly unstructured, yet potentially useful, temporal information about patient's evolution.

2.2 Data preparation

To perform the analysis, a thorough data preparation stage was carried out, involving several imputation methods of missing data to complete synchronous, but variable in length, time series per patient. In particular, we used the previous day value for the same patient, or zero if no measurements were available. For information about ICU stay we created an increasing integer variable, for each day the patient stayed in ICU. For clinical values, missing data was replaced with normal reference values. The rest of missing data were replaced with zeros. To account for the temporal evolution, we considered both the original table, with one record per patient, used for t-SNE projection according to time series similarities, and an *unpivoted* version of it containing $\mathsf{patient-day}$ records used for visualization. Selected treatments were included in the records using one-hot encoding in a per-day basis.

2.3 Morphing projections for interactive rearrangement of data

As a main interactive visualization approach, we used the *morphing projections* (MP) technique [1]. MP is based on linearly mixing two or more meaningful views, resulting into smooth transitions among different but complementary

[1] `https://www.hmhospitales.com/coronavirus/covid-data-save-lives`

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

representations of data and knowledge. In its most common form it is applied to 2D scatterplot views. MP assumes p *basis views* –or *encodings*–, each composed of n points, $\{\mathbf{p}_i^j\}_{i=1\ldots n}$, where $j = 1 \ldots p$, and p user-steerable *mixing coefficients* $\lambda_1 \ldots \lambda_p$ where $\sum_i \lambda_i = 1$. Each encoding is supposed to contain a meaningful –typically 2D or 3D– arrangement of the samples, that may include supervised encodings (e.g. circular, linear, matrix, geolocalized arrangements of the data) or unsupervised ones, such as 2D DR projections of the samples according to similarities using some metric. During data exploration, the user steers the weights λ_i and the interface renders a *live* scatterplot:

$$\mathbf{p}_i(t) = \sum_j \lambda_j(t)\mathbf{p}_i^j \tag{1}$$

In our setting, we considered user-configurable basic *supervised* encodings, including, two vertical positions for sex $[1, 0]$, $[-1, 0]$, linear (e.g. vertical) positions $\{[0, k]\}$, for $k = 1 \ldots N$ timestamps, and circular $\{[\cos(2\pi k/M), \sin(2\pi k/M)]\}$ with $k = 1 \ldots M$, for other attributes like patients. Note that combining these encodings through (1) yields meaningful composed encodings, such as sets of circles (e.g. two circles for male and female patients) or matrix encodings (linear $\times$ vertical encodings).

Also, *unsupervised* encodings were included using the t-SNE algorithm [2] taking similarities between all patients' time series. Such approach provides a 2D arrangement of the patients according to similarities in the evolution of their health state, including vital signs, where patients sharing similar conditions are grouped in specific regions of the map identifiable after some data exploration.

2.4 Dealing with time series

The time series in this example are very challenging to deal with. They include, not only missing data, but also different time frames (some patients progress more rapidly than others). Thus, the quality of any DR technique is severely affected by the employed similarity metric. In this paper we explore the well-known *dynamic time-warping* (DTW), first defined by Sakoe and Chiba [3] in 1978 for spoken word recognition, which has long ago proven to be superior to the Euclidean distance and applied in many different fields [4], including medicine and bioinformatics –see, for instance, [5].

We chose five variables commonly monitored during patient stay: heart rate (FC), leukocytes (LEU), lymphocytes (LIN), platelets (PLAQ) and partial pressure of Oxygen (PO2). Similarities for each variable are computed using the DTW method and normalized, obtaining five different distance matrices $d_{ij}^1, \ldots, d_{ij}^5$. The final distance matrix is calculated as $\sqrt{\sum_k (d_{ij}^k)^2}$, and is used for training a t-SNE model. With this metric, the resulting projection conveys information on the overall similarity in the time series of the five variables, resulting in a layout that reflects the casuistic of clinical evolution patterns.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

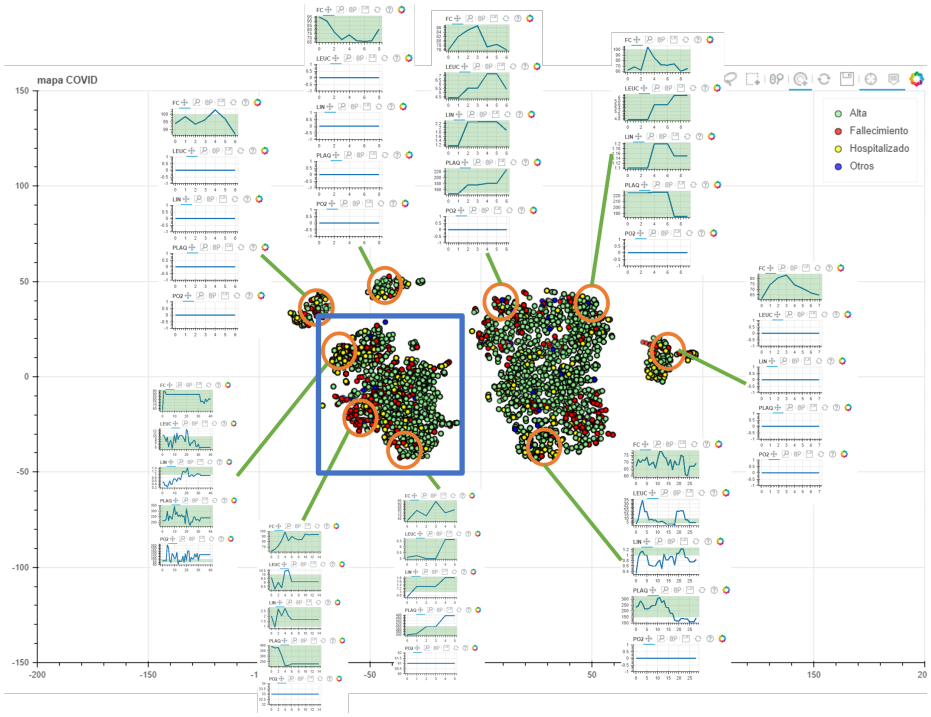

Fig. 1: t-SNE trained with the time series of FC, LEU, LIN, PLAQ, PO2, described in 2.4. Each point contains, collapsed, all the time steps for a patient.

3 Results and Discussion

Fig. 1 shows a t-SNE trained with the time series data. This analysis revealed 5 clusters, 4 of them including patients with missing data, and one cluster (blue box) containing all patients with valid data for all variables of analysis.

Stratification of patients' data is done in an intuitive, fast and visually rich way using MP. By moving a slider, the expert user can steer the weight $\lambda_i(t)$ of each encoding, and smoothly transition between different visual arrangements; for example, segregating the data by the final outcome of the patient (dead, in hospital, discharged or other). Furthermore, temporal information can be incorporated, again in an easy way, by including it in an appropriate encoding (e.g. vertical). By moving a slider, the user can disaggregate patient data and temporal records appear as vertically arranged data points, each one corresponding to a record in the $\mathsf{patient} - \mathsf{day}$ table.

A full example is shown in Fig. 2, where the DAYS-IN-ICU variable has been encoded in a vertical arrangement. The figure also highlights the power of interactive analysis using MP as a tool for data exploration and hypothesis generation. Several encodings are activated sequentially by the user during the analysis, to show the different evolution of patients according to sex, which was

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

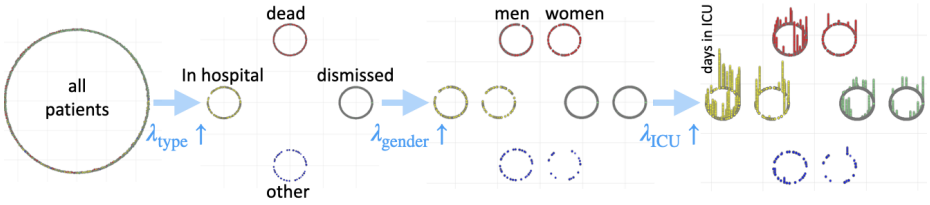

Fig. 2: Segregating data by patient status, sex and incorporating days in ICU in a vertical encoding shows how men have a strong tendency to stay in ICU for longer than women for those with fatal results or those who are still hospitalized, but is more similar for patients who were discharged. This is consistent with the better evolution of women seen during the first wave.

something observed during the first wave of the COVID-19 pandemic. The user starts with all the patients arranged in a circle, and may incorporate different encodings in a highly interactive exploratory process which provides immediate feedback. Thus, the visual display of the groups resulting from stratification reveals differences in the number of items for each group that can be pre-attentively spotted by quick visual queries.

Fig. 3 presents another analysis pipeline, taking into account the clinical evolution of the patients, that shows how this technique greatly enriches the information contained in the t-SNE plot. By adding the temporal information, the user can promptly see that the patients that spent longer in ICU lay only in certain areas of the t-SNE, which means they share a similar evolution.

The MP technique also allowed a rapid and global assessment of the impact of drug administration on any patient variable. For example, Dexamethasone was observed to have a positive effect on the patient's PO2, unlike other drugs such as Aminoquinolines and Glucocorticoids (see demo video).

4 Conclusions

Interactive visual analytics using MP allows the user to explore the data by combining different criteria, such as sex, the value of a clinical variable, medication, or even machine learning powered layouts, such as DR projections.

We included a temporal analysis establishing the $\mathsf{patient} - \mathsf{day}$ as the basic element, and using *dynamic time warping* as a similarity measure of the patient progress for the t-SNE model, so that patients with similar clinical evolution are grouped together. This temporal information is naturally integrated into the interactive analysis by means of the MP technique using specialized encodings.

The proposed approach has been demonstrated through an application example involving highly unstructured and multifaceted biomedical data, including clinical and medication information that was sampled asynchronously. This method is highly interactive and visual, thus allowing to exploit the soft and fault-tolerant nature of the human cognitive process. In addition, the MP tech-

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

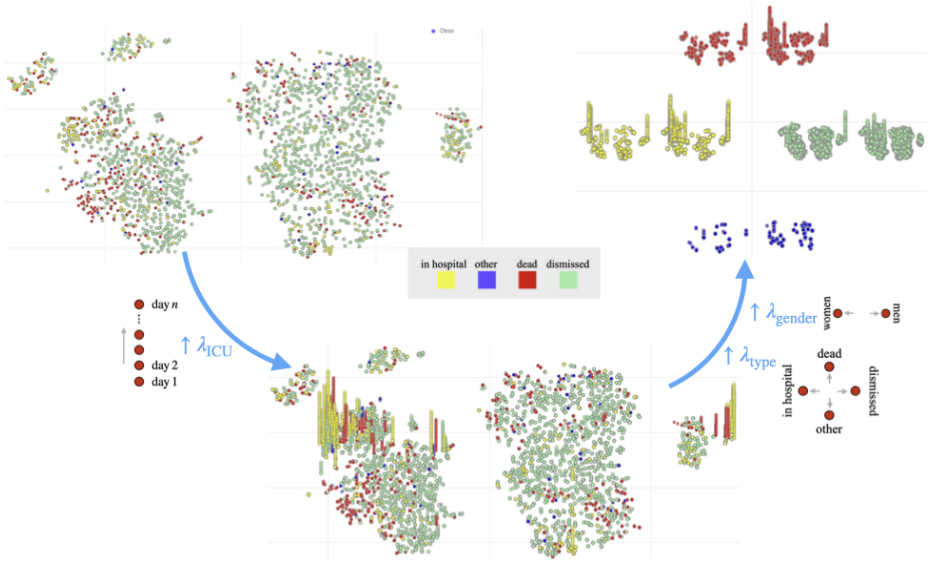

Fig. 3: By including the temporal data as a vertical encoding, the days spent in ICU for each patient appear as vertical bars over the t-SNE, clustered in specific regions. More detailed analysis over the patient's evolution in time can be easily performed after segregation of data, as in the previous example.

nique allows the user to maintain a stable visual trace of the data between views representing different knowledge domains. These unique features enable the expert user to gain a broad view, acquire new knowledge, and formulate new hypotheses in a highly unstructured multi-domain data scenario.

References

[1] Ignacio Díaz, José M Enguita, Ana González, Diego García, Abel A Cuadrado, María D Chiara, and Nuria Valdés. Morphing projections: a new visual technique for fast and interactive large-scale analysis of biomedical datasets. *Bioinformatics*, 37(11):1571–1580, 2021.

[2] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[3] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[4] Chotirat Ann Ratanamahatana and Eamonn Keogh. *Making Time-series Classification More Accurate Using Learned Constraints*, pages 11–22.

[5] Aach J. and Church G.M. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495 – 508, 2001. Cited by: 360; All Open Access, Bronze Open Access.