Machine Learning For Exploratory Data Analysis And Model Diagnosis In Oncology

PumasAI

Lucas Pereira[1], Mohamed Tarek Mohamed[1], Lorenzo Contento[1]

1. Pumas-AI Inc. Dover, DE 19901

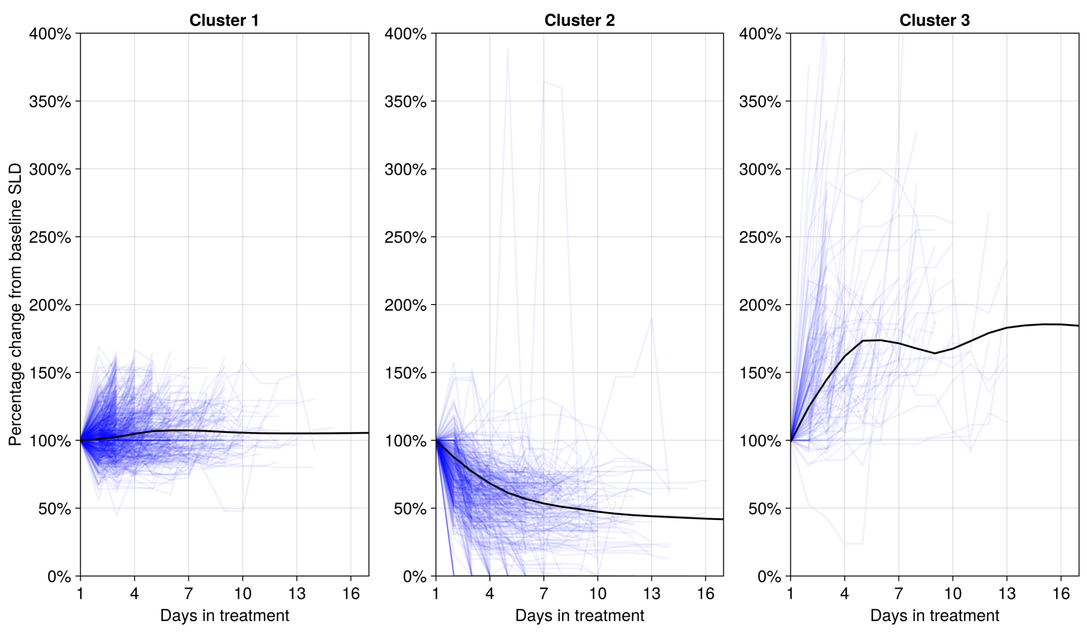
INTRODUCTION

Visualizing large longitudinal datasets is not a trivial task. We will present an application of machine learning (ML) to cluster longitudinal irregular data. **Clustering** is an unsupervised ML task which can aid in **exploratory data analysis**. Clustering can ease visualization of the various patterns in the data. It enables automatic identification of subpopulations in the response (e.g. responders and non-responders). And it can help identify potential outliers. We also demonstrate how clusters can be used to **stratify model diagnostics** by subpopulations to understand which ones are fitted better or worse, augmenting the model development process.

As an example, we use the dataset from [1]. The authors collated data from five clinical trials investigating Atezolizumab, an immune checkpoint inhibitor. In total, 1472 patients with at least three measurements of the diameter of the target lesion are included, 652 of whom have at least six data points. The authors also investigated the performance of the following **classical oncology models** in real-word data: exponential, logistic, classic Bertalanffy, general Bertalanffy, classic Gompertz, and general Gompertz. We focus on patients with 6 or more measurements; and on the general Bertalanffy model (since it had one of the best performances in the original study).

RESULTS

As shown in the following figure, the combination of DTW and K-Medoids was able to **split the trends** in three visually distinct groups.



And, from left to right, they could be labeled "fluctuate", "down" and "up". However, it is important to point out that K-Medoids doesn't project any meaning or interpretation onto the clusters. It is up to the researcher to make sense of the groupings and fit them into the bigger picture of the analyses and context. In fact, even the number of clusters (here, 3), is usually experimented with, to learn what is the best number of groups that describes the data or just explore more specific groupings using more clusters.

For the next step, we proceeded to build an NLME model based on the **general Bertalanffy** model. The formulation of its analytical solution is shown in the following equation:

$$v(t) = v_\infty \left\{ 1 + \left[\left(\frac{v_0}{v_\infty} \right)^{1/3} - 1 \right] e^{-\omega t} \right\}^3$$

v_∞ is the steady-state lesion volume by the end of the trend; t is time; v_0 is the baseline volume; and ω is a dimensional parameter resulting from the reformulation in [2]. The NLME model was kept simple, since the goal is to showcase the use of population stratification in PKPD analyses, and not develop a state-of-the-art model. An additive residual error term, alongside random effects on the parameters (v_∞, v_0, ω), was included.

CONTACT

www.pumas.ai/resources

Lucas Pereira lucas.p@pumas.ai

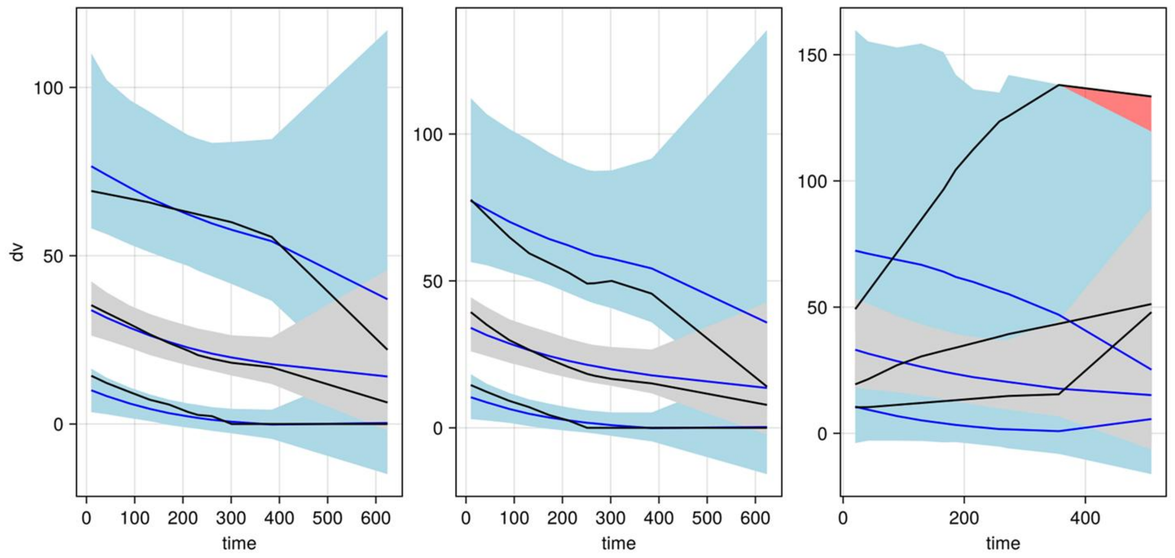
METHODS

Initially, a combination of DTW and K-Medoids enabled the clustering of the length-varying tumor data from [1]. **DTW (dynamic time warping)** is a method to calculate the dissimilarity between two sequences of numbers, even if they have different lengths and uneven spacing between elements. When comparing two sequences, the algorithm maps every element of one to at least one element in the other. Therefore, DTW can be used to process trends of measurements of sum of largest lesion diameters, a common observation in oncology used to keep track of treatment response.

Additionally, the total cost associated with the mapping is determined by DTW. And **K-Medoids**, a clustering algorithm, can use these pairwise costs as distances between series of tumor measurements. The population was clustered into 3 groups: trends that decrease (“down”), increase (“up”), or are stable (“fluctuate”). Then, a new set of patients was sampled from the clusters, containing 10 subjects from “up” and 40 from “down”. And a nonlinear mixed-effects (NLME) model was fitted to this new group. The dynamics of this model were based on a reformulation of the classic general Bertalanffy [2].

RESULTS

With the clusters and the model at hand, we created a new (“mix”) population by sampling 10 subjects from the “up” cluster and 40 subjects from the “down” cluster. After fitting the model with first-order conditional estimate (FOCE), it was used to simulate observations for 1000 populations. This enables us to use visual predictive check (VPC) plots, as shown below. From left to right, the VPCs refer to the populations “mix”, “down” and “up”, respectively.



VPCs are a model diagnostic tool used to test if model predictions aren't too far from the data. The three bands represent quantiles from the simulations, specifically 90th, 50th and 10th quantiles from top to bottom. The blue lines are the medians of each band; and the black lines are the analogous for the data. As shown on the left, the model has a reasonable alignment with the entire new population of 50 subjects. A similar behavior is seen on the center plot. However, on the right, a poor fit is indicated by the discrepancies between each pair of blue and black lines; and by the red region at the top right, referring to outliers.

CONCLUSIONS

The described experiments show the usefulness of clustering for population stratification. With no intervention, it **spotted the existence of relevant subgroups**, which usually isn't trivial. Furthermore, clustering algorithms are application-agnostic; don't impose constraints on the data; and don't introduce bias in the interpretation. Additionally, the subpopulations identified enabled more experiments, including the **stratification of model diagnostics**, studying the quality of the fit for each group. Ultimately, clustering and DTW were shown to be useful tools for exploratory data analysis and model diagnostics.

REFERENCES

1. N. Ghaffari Laleh, et al. Classical mathematical models for prediction of response to chemotherapy and immunotherapy. PLOS Computational Biology, 18(2):1–18, 2022.
2. A. D. Blaom and S. Okon. New tools for comparing classical and neural ODE models for tumor growth, 2025.