

OptiClust4Rec: Unsupervised Data-Driven Methodology for Quality of Life Recommendations During a Medical Therapy (Extended Abstract)

Juba Agoun^{1,*}, Yanis Bouallouche² and Mohand-Saïd Hacid¹

¹Université Claude Bernard Lyon 1, CNRS, LIRIS, UMR5205, 69100 Villeurbanne, France

²Université de Montpellier, 34095 Montpellier, France

Abstract

Upon the introduction of novel medical therapies, an array of semantically different data is gathered from the participant's cohort. Unsupervised learning is always privileged as a preliminary step for data investigation, to extract valuable information before embarking on the tedious task of data labeling. Clustering is one of the techniques that provide a comprehensive overview for exploratory data analysis, aiding in the identification of patient communities. With OptClust4Rec, we provide a characterization of clusters, from which we can derive recommendations for patients undergoing therapy treatment. Our focus is on optimizing the clustering and the dimensionality reduction based on concise metrics and data topology analysis.

Keywords

Clustering, Optimization, Data analysis, Unsupervised learning

1. Introduction

The healthcare sector generates a substantial amount of data daily. New treatment therapies are tested on patients during clinical trials. The acquired data are semantically different, covering health status, side effects, work potential, and lifestyle. It is collected to unveil causal inferences regarding treatment effectiveness. These data have therefore become predicting subsequent future health conditions, reducing the cost of treatments, and improving the quality of life in general. For instance, to improve patients' immunotherapy treatment, the analysis of upstream medical data enables the production of recommendations and guidelines for practitioners.

In aiding medical decision-making, classical patient clustering proves to be a dependable approach. This involves identifying dominant characteristics within patient groups and tracking their health progression. Unsupervised techniques allow for an initial analysis of data relationships, without necessitating specialized domain knowledge. While supervised methods have shown their efficacy, building labeled datasets remains a time-consuming and labor-intensive

Proceedings of the Demonstration Track at International Conference on Cooperative Information Systems 2023, CoopIS 2023, Groningen, The Netherlands, October 30 - November 3, 2023


*Corresponding author.

✉ juba.agoun@univ-lyon1.fr (J. Agoun); yanis.bouallouche@etu.umontpellier.fr (Y. Bouallouche); mohand-said.hacid@univ-lyon1.fr (M. Hacid)

ORCID 0000-0003-4911-0617 (J. Agoun)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

task. Hence, with OptiClust4Rec, we introduce an unsupervised data-driven methodology, enabling the discovery of concealed patterns in clinical patient data alongside questionnaire responses.

OptiClust4Rec¹ (Optimized Clustering for recommendation) introduces a user-friendly web-based interface aimed at assessing clustering algorithm capabilities for specific datasets. This evaluation employs statistical measures based on internal metrics, complemented by visual exploration to analyze measure variations with diverse parameters and dimensionality reduction techniques. The tool guides users through optimizing the clustering process, producing distinct clusters, and then characterizing them by extracting essential features. To generate patient recommendations, we utilize two datasets: one with patient analyses and information, and another containing their questionnaire responses. OptiClust4Rec is designed to cluster patients based on both sets of data and label the clusters, which, consequently, provides end-users with association of rules and the subsequent production of recommendations.

2. System overview

In our approach, we consider two challenges. First, we propose a set of metrics and visualization tools that will enable users to optimize data clustering. Second, leveraging the clustering results, we will focus on each cluster to extract the variables that characterize it, essentially creating a form of automatic labeling based on what is known as Salient Features.

2.1. Clustering

There are many well-understood techniques to draw upon; Centroid-based, Connectivity-based, Grid-based, and Density-based. Choosing the right method with the right parameters for a given dataset is performed following any deterministic method. In practice, to choose between clustering methods, or to determine the number of clusters as input value, required testing different algorithms. Indeed, researchers either select a default method (*e.g.*, *k*-means) with a number of clusters well-known depending on domain knowledge, or subjectively choose the most recent method available [1]. Next, we will introduce the two optimizations we target with our tool.

2.1.1. Find well-adapted clustering algorithm

Among the existing clustering algorithms, the challenge is to identify the appropriate model for the data being analyzed. Our idea relies on the adoption of Persistent Homology, a fundamental technique within the domain of Topological Data Analysis (TDA) [2]. This approach systematically examines the relationships among data points across various scales, providing invaluable insights into the presence, geometric arrangement, and density of potential clusters. The application of Persistent Homology results in a persistent diagram.

In the persistent diagram, each distinct color represents a unique homology group. Specifically, H_0 represents the connected components, acting as a guide in estimating a potential cluster

¹Visual summary of our approach through this illustration <https://tinyurl.com/OptiClust4Rec-Figures>

count. H1 points denote one-dimensional voids, while the H2 group outlines two-dimensional voids, often referred to as *cavities*. The diagram intricately traces the horological changes as we explore various scales of distance or proximity within the dataset. Each connected component and void materializes as a point on the graph, with its inception (appearance) and cessation (disappearance) plotted along the horizontal and vertical axes. A significant deviation of a point from the diagonal indicates the enduring presence of the corresponding component, potentially signifying robust and lasting structures within the data. Upon thorough examination of the persistent diagram for a given dataset, the notable deviation of certain H1 and H2 points from the diagonal strongly suggests the presence of nonlinear structures. These may involve spherical forms or overlapping clusters. Consequently, this observation strongly supports the recommendation of employing a density-based algorithm like DBSCAN, renowned for its capability to perform exceptionally well in complex scenarios of this nature.

2.1.2. Find optimal k number of cluster

Non-density-based clustering techniques require as an input the number of clusters to be shaped. Thus, determining the exact or optimum number is a challenging task due to the absence of a universal method for identifying the ideal number for a given dataset. The elbow method and the Silhouette score are commonly used methods to find the optimal number of clusters.

The elbow method assesses cluster compactness by computing the Within-Cluster Sum of Squared (WCSS) for different cluster numbers, observing a decrease in WCSS as k increases, indicating improved compactness. However, there comes a point where adding more clusters no longer enhances quality. The point of this diminishing return, visually depicted as an *elbow* in the WCSS evolution chart with respect to cluster numbers, indicates the optimal cluster count. Nevertheless, it's important to note that this method does not account for cluster separation, an important aspect of ideal clustering.

In contrast, the silhouette score provides a more comprehensive evaluation of cluster quality by considering both cohesion (average distance within a cluster) and separation (average distance to the nearest neighboring cluster) for each data point. However, it may encounter challenges when identifying complex clusters with diverse shapes. It's important to note that in some instances, clusters identified by the silhouette score may exhibit uneven distributions, potentially resulting in clusters with only a few observations while the rest are dispersed across other clusters.

We proceed to determine the optimal value of k by using two internal metrics: connectivity and variability [3]. The goal is to examine how variability changes with respect to connectivity. With this pattern analysis, we aim to identify the characteristic point where variability decreases significantly and connectivity slightly increases. This point is visualized as a *knee* on the graph depicting the evolution of variability with respect to connectivity, and it signifies the optimal number of clusters based on [4]. Our approach is applied to seven different algorithms and is rounded off with a voting mechanism. For example, if four out of the seven algorithms advocate for four clusters as the ideal number, then four clusters will be considered optimal.

Dealing with high-dimensional data presents challenges, emphasizing the need for dimensionality reduction. The *curse of dimensionality* poses a significant problem, leading us to favor UMAP over Principal Component Analysis (PCA) for its efficiency in preserving data structure

and topology. We explore various dimensionality reductions, considering a maximum of \sqrt{N} dimensions for N observations [5]. For each reduction, we search for the optimal number of clusters and examine variance in cluster numbers across dimensions, using lower variance as an indicator of particular interest and influencing our selection of cluster count.

2.2. Characterizing clusters

Given our primary objective of working with semantically various data to derive association rules, it is important that our data be appropriately labeled. Following the application of clustering, our aim is to distill the prominent characteristics of each cluster by extracting the salient features. To accomplish this, we adopt the approach outlined in [6]. In this step, the observations are categorized into in-pattern and out-pattern records. With the analysis of in-patterns and out-patterns within a given cluster, we can pinpoint the salient features and discern whether the associated variables exhibit high or low values.

3. OptiClust4Rec



Figure 1: Labeled screenshot of OptiClust4Rec interface.

OptiClust4Rec² is a web-based application designed for the analysis and visualization of both medical and nonmedical datasets. It primarily employs unsupervised techniques, mainly clustering, and offers guidance through the utilization of dimensionality reduction, topological data analysis, and automated labeling methods. As the screenshots in figure 1 illustrate, OptiClust4Rec has several user interfaces, mainly three, which we detail in the following sections:

- A. The area presents the results of different clustering operations in different dimensions. After loading the datasets in another interface, the user obtains results compared with the most well-known internal metrics of the literature.
- B. The area displays the result of the persistence homology, which provides information on the presence of cavities. The more H1 and H2 points we find deviating from the diagonal, the more we recommend the use of density-based methods.
- C. The area is intended to display the results of cluster characterization. Once the user selects the appropriate clustering method, each cluster is labeled with the salient variables.

Given the result of each dataset cluster characterization, the user can derive correlations between the clusters of the different semantic datasets.

Acknowledgments

This research is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 875171, project QUALITOP (Monitoring multidimensional aspects of Quality of Life after cancer ImmunoTherapy - an Open smart digital Platform for personalized prevention and patient management).

References

- [1] A. J. Parker, A. S. Barnard, Selecting appropriate clustering methods for materials science applications of machine learning, *Advanced Theory and Simulations* 2 (2019) 1900145.
- [2] L. Wasserman, Topological data analysis, *Annual Review of Statistics and Its Application* 5 (2018) 501–532.
- [3] J. Handl, J. Knowles, D. Kell, Bioinformatics computational cluster validation in post-genomic data analysis, *Bioinformatics (Oxford, England)* 21 (2005) 3201–12.
- [4] V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan, Finding a "kneedle" in a haystack: Detecting knee points in system behavior, in: *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.
- [5] J. Hua, Z. Xiong, J. Lowey, E. Suh, E. Dougherty, Optimal number of features as a function of sample size for various classification rules, *Bioinformatics (Oxford, England)* 21 (2005) 1509–15.
- [6] M. R. Khoie, T. Tabrizi, E. Khorasani, N. Marhamati, A hospital recommendation system based on patient satisfaction survey, *Applied Sciences* 7 (2017) 966.

²Link to tool and video: <https://tinyurl.com/Demo-OptiClust4Rec>