Categories and Clusters to investigate Similarities in Diabetic Kidney Disease Patients

Categorie e clusters per investigare la similarità fra i pazienti affetti da nefropatia diabetica

Veronica Distefano, Maria Mannone, Claudio Silvestri, and Irene Poli

Abstract Heterogeneous responses to therapeutical treatments across patients and over time is a common and serious problem for several diseases. Precision medicine research focuses in developing procedures to take treatment decisions for the individual patient using all the information available for the patient, including demographic and clinical variables and the response to the followed treatment. In this paper we adopt category theory and the cluster analysis to achieve insight into specific disease pathways and patient subgroups. We analyze a longitudinal dataset of patients affected by diabetic kidney disease (highly prevalent in type 2 diabetes) and monitored at different time points in the response to various treatment regimes. This analysis, based on distances between patients in different time points and in time evolution, divides patients into clusters that show the relevant role of some variables in affecting the progress of the disease.

Abstract *L'eterogeneità nella risposta a trattamenti terapeutici tra pazienti e nella sua evoluzione temporale rappresenta un problema comune a molte malattie. La medicina di precisione si propone di sviluppare metodologie di supporto alle decisioni di trattamento per ogni singolo paziente usando tutte le informazioni disponibili sul paziente, includendo perciò le variabili demografiche, le variabili cliniche e la risposta ai trattamenti. In questo lavoro noi adottiamo la teoria delle categorie e la cluster analysis per ottenere elementi informativi su diversi sviluppi della*

Veronica Distefano
European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: veronica.distefano@unive.it

Maria Mannone
European Centre for Living Technology, Ca' Foscari University of Venice, Italy, and Department of Mathematics and Computer Sciences, University of Palermo, Italy, e-mail: maria.mannone@unive.it,mariacaterina.mannone@unipa.it

Claudio Silvestri
Dipartimento di Scienze Ambientali, Informatica e Statistica and European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: claudio.silvestri@unive.it

Irene Poli
European Centre for Living Technology, Ca' Foscari University of Venice, Italy, e-mail: irenpoli@unive.it

malattia nel tempo e sull'esistenza di gruppi di pazienti con comportamenti diversi. Analizziamo un insieme di dati longitudinali relativi a pazienti con diabete di tipo 2 e complicazioni renali, osservati in differenti punti temporali con riferimento a un insieme di variabili e alla risposta a differenti trattamenti. Questa analisi, basata sulle distanze tra pazienti in successivi punti temporali, individua clusters di pazienti e il ruolo di certe variabili nell'influenzare la progressione della malattia.

Key words: category theory, cluster analysis, DKD disease

1 Introduction

Precision medicine greatly profits from statistical and mathematical techniques to envisage differences and similarities between patients, their treatments, and outcomes. The increasing interest on these topics leads to a more extended development and use of methodological procedures. One of the research areas where precision medicine is currently considered is the treatment of diabetes of type 2, with kidney clinical complication (DKD). DKD patients show significative heterogeneity in the disease progress, and thus there is a clinical need for individualized treatments. Patient clustering [1, 6] appears as a useful explorative way to achieve some information about the evolution of the disease. We focus on a longitudinal dataset of DKD patients from the DC-ren project,[1] and we aim to highlight similarities between patients, in terms of initial conditions, therapeutical treatments, and responses to the treatments. The dataset consists of mixed data, which include quantitative and qualitative variables, concerning clinical-laboratorial data, socio-demographical aspects, and different treatment responses.
In this paper, we aim to derive clusters of patients observed in different time points, where the clustering approach will be based on distances between patients. The evaluation of distance between patients is an essential step to investigate patients' characteristic profiles. In order to derive an integrated approach to visualize and highlight dynamic patterns of the disease, we adopt the category theory, an abstract branch of mathematics, developed to formalize the concept of *transformations between transformations* in a flexible way [4]. This approach is used in a variety of areas of research, which include biology, physics, chemistry, and computer science [2], and it is useful to investigate problems in an abstract way and visualize connections and temporal dynamics. In this framework, a category is constituted by objects (points) and morphisms between them (arrows), and provides a clear way to model similarity and equivalence. One of the most powerful ideas of category theory is the notion of *functor*, which can be thought as a generalization of the concept of function. A functor maps objects and morphisms from a category to objects and morphisms of another category. A mapping between functors is a natural transformation, and leads to generate nested structures. The novelty of this work consists in deriving an integrative approach to visualize and explore patterns in longitudinal

[1] https://dc-ren.eu/

data. Longitudinal data on DKD are collected in visits at baseline and in subsequent follow-ups. We focus on the first three time points t_0, t_1, t_2, comparing the response to the therapy treatment of the set of patients. In order to evaluate the heterogeneity of patients at different time points and how this heterogeneity evolves in time, we evaluate matrices of distances between patients. We introduce $D(t_0)$ as the matrix of distances between patients at time t_0; $D(t_1)$ and $D(t_2)$ as the matrices of distances between patients at t_1 and t_2, respectively. We also introduce $D(t_0,t_1)$ as the matrix of distances between $D(t_0), D(t_1)$, and $D(t_1,t_2)$ as the matrix of distances between $D(t_1), D(t_2)$. With this approach adopting category theory and cluster analysis, we achieve a small set of clusters, which represent the most similar patients in their behavior in time and response to the therapy treatments. These results will be very helpful in building a decision system which allows to derive the best treatment for each patient. The paper is organized as follows. In Section 2, we adopt elements of the category theory to envisage a strategy for deriving matrices of distance between patients, and in Section 3, we build clusters of patients with respect to the progression of the disease.

2 Method

In our analysis, we will consider a set of data concerning n patients, p variables, and three time points t_0, t_1, t_2. Each element is the observation $x_i^j(t_k)$, where: i indicates the individual (the patient), $i = 1, ..., n$; j indicates the variable (X^j) on which we observe x_i^j, $j = 1, ..., p$; k indicates the time point, $k = 0, 1, 2$, and thus there are three time points: t_0 (also called *baseline*), and t_1, t_2 (the first and second *follow-ups*, respectively). We define two kinds of distances: distance $d^j_{i,i'}(t_k,t_k)$ between observations of variable j at the time k for different patients $i, i' = 1, ..., n$ (horizontal distance); distance $d^j_{i,i}(t_k,t_{k'})$ between observations of variable j through different times $k, k' = 0, 1, 2$ for patient i (vertical distance). Having built a set of distance values, we can achieve an *enriched category* with metrics in $\mathbb{R}$ [4]. More precisely, we can describe observations and distances as an enriched *double category* whose objects are $x_i^j(t_k)$ and whose morphisms are vertical and horizontal distances, as in the following diagram 1.

$$\begin{tikzcd}
x_i^j(t_0) \arrow[r, "d^j_{ii'}(t_0,t_0)"] \arrow[d, "d^j_{ii}(t_0,t_1)"'] & x_{i'}^j(t_0) \arrow[d, "d^j_{i'i'}(t_0,t_1)"'] \\
x_i^j(t_1) \arrow[r, "d^j_{ii'}(t_1,t_1)"] & x_{i'}^j(t_1)
\end{tikzcd} \tag{1}$$

Observations for the same variable through time and patients constitute a lattice. Horizontal composition shows comparisons between multiple patients at the same time; vertical composition shows comparisons of the same patient through time. Mappings from variables to variables can be formalized as *functors*. There is a lattice for each variable. These lattices are the vertical sections in the representation of Figure 1 (left). In fact, observations and distances for each patient are represented by transversal sections (Figure 1, right). Functors map a lattice into the other. The outcomes (success/unsuccess) of the therapeutical treatment are evaluated through the variations of a response variable. We compute the dissimilarity matrices, to compare patients and their disease time evolution. We evaluate as a first step the information in the variable values of each patient at certain times. The i-th patient $p_i(t)$ is described by the vector $[x_i^1(t_k), x_i^2(t_k), ..., x_i^p(t_k)]$. Each element of the dissimilarity matrix is the distance between patients, as described in Figure 2.

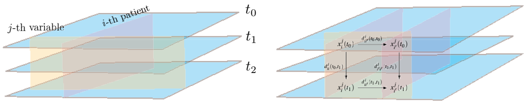

Fig. 1 Representation of the dataset.

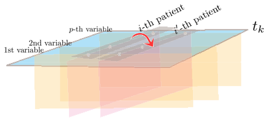

Fig. 2 Representation of matrix elements of $D(t_k)$.

To measure the distance, we adopt the coefficient $s(p_i, p_{i'})$ proposed in [5], which takes the following expression: $s(p_i, p_{i'}) = \frac{\sum_{j=1}^{p} s^j(x_i^j, x_{i'}^j)\delta^j(x_i^j, x_{i'}^j)}{\sum_{j=1}^{p} \delta^j(x_i^j, x_{i'}^j)}$, where: p_i is the i-th patient, $p_{i'}$ is the i'-th patient, s^j is based on the Gower similarity [3], x_i^j is the value of the j-th variable for the i-th patient, $x_{i'}^j$ is the value of the j-th variable for the i'-th patient; $\delta^j(x_i^j, x_{i'}^j)$ is a coefficient to be 0 if p_i or $p_{i'}$ have a missing value for the j-th variable, and 1 if they do not. The elements of the dissimilarity matrix, describing the *distance between patients*, are computed as $d(p_i(t_k), p_{i'}(t_k)) = 1 - s(p_i(t_k), p_{i'}(t_k))$. The dissimilarity matrices, computed at different time points, are then $D(t_0), D(t_1), D(t_2)$, and on these matrices we build matrices of distances of distances $D(t_0, t_1), D(t_1, t_2)$.

3 Results

In order to evaluate the behaviors of patients with respect to variables and therapeutical treatments, we build clusters of patients according to the measures of distance described in Section 2. The dataset, used in the DC-ren project and based on the PROVALID study,[2] is about diabetic kidney disease (DKD), with $n = 241$ patients observed in longitudinal way in three data points and $p = 21$ variables, which include clinical and social-demographic variables and treatment responses. In order to find clusters of patients, we build matrices of distances between patients at the three data points, $D(t_0)$, $D(t_1)$, $D(t_2)$, and, based on them, matrices of distances between distances, $D(t_0, t_1)$, $D(t_1, t_2)$. The data are collected at baseline (t_0) and follow-ups

[2] SysKid project `https://cordis.europa.eu/project/id/241544/`

(t_1, t_2), and contain information on the therapy adopted. The patients received four different treatments, a_1, a_2, a_3, a_4. The outcome is a binary variable, indicating *success* or *unsuccess* of the therapeutical treatment. In this paper, we adopt the hierarchical clustering method without deciding *a priori* the number of clusters.

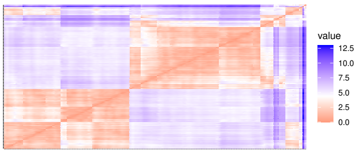

Fig. 3 Matrix $D(t_0,t_1)$. (Darker blue indicates a greater dissimilarity).

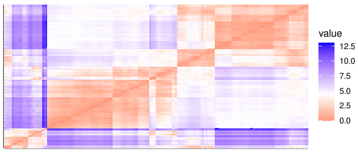

Fig. 4 $D(t_1,t_2)$. (Darker blue indicates a greater dissimilarity).

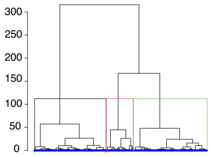

Fig. 5 Dendrogram of $D(t_0,t_1)$.

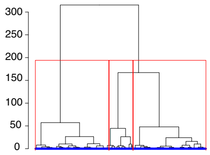

Fig. 6 Dendrogram of $D(t_1,t_2)$.

To obtain $D(t_0,t_1)$, represented in Figure 3, we compute $D(t_0)$ associated with the distances between patients at t_0, and matrix $D(t_1)$ associated with the distances between patients at t_1, both evaluated with Gower distance. To obtain $D(t_1,t_2)$, represented in Figure 4, we follow the same procedure. The sequence of clusters obtained by using a hierarchical clustering is visualized through the Ward dendrogram by using the matrices $D(t_0,t_1)$ and $D(t_1,t_2)$, and we achieve K = 3 as the optimal number of clusters. The height of the dendrogram is the distance between the clusters, as shown in Figures 5 and 6. In our study, the comparison between four linkage methods (average, single, complete, ward), shows that the Ward-type linkage method identifies the better clustering structure. The Ward method finds the distance between two clusters as the minimum within-cluster variance. In Table 1, we can see that the mean values of the vast majority of the variables within clusters 2 and 3 in $D(t_0,t_1)$ and the variables within clusters 1 and 3 in $D(t_1,t_2)$ do not show significant differences. On the contrary, we notice relevant differences in cluster 1 in $D(t_0,t_1)$ and cluster 2 in $D(t_1,t_2)$. In particular, the patients in these clusters present higher body mass index (BMI) and triglyceride values, while the levels of HbA1c, cholesterol HDL, and estimated Glomerular Filtration Rate (eGFR) are lower. In particular, the mean eGFR of these patients is <60 mL/min/1.73 m^2, which increases the risk factor to cardiovascular disease. Figure 7 shows the boxplots of clusters of patients for different treatments and different eGFR levels. The 47.3% of patients in cluster 1 in $D(t_0,t_1)$, is present in cluster 2 in $D(t_1,t_2)$. These patients are mostly treated with a_1 and a_3. The 66% of them had *unsuccessful* outcome in $D(t_0,t_1)$ and $D(t_1,t_2)$. From this analysis, it is possible to find groups of patients with behaviors different between them, and with respect to the time evolution of the disease.

Table 1 Description of the clusters of distances between patients: $D(t_0,t_1)$ and $D(t_1,t_2)$, with variable mean values and standard deviation values within within brackets.

	$D(t_0,t_1)$			$D(t_1,t_2)$		
Variables	cluster1 (n= 38)	cluster2 (n= 100)	cluster3 (n= 103)	cluster 1 (n= 103)	cluster 2 (n= 34)	cluster 3 (n= 104)
eGFR	55.13 (± 21.59)	65.65 (± 18.34)	62.73 (± 17.71)	62.14 (± 17.88)	57.24 (± 21.90)	59.25 (± 17.27)
Age	66.55 (± 11.38)	69.55 (± 7.81)	66.10 (± 8.35)	68.03 (± 7.76)	65.82 (± 11.52)	65.38 (± 8.60)
BMI	34.74 (± 9.74)	30.87 (± 6.52)	30.27 (± 5.81)	31.24 (± 6.76)	33.90 (± 10.45)	30.10 (± 5.98)
Body weight	96.68 (± 22.35)	85.43 (± 16.35)	85.32 (± 13.97)	86.38 (± 17.00)	94.62 (± 24.09)	84.88 (± 14.05)
Systolic	137.29 (± 15.80)	137.08 (± 14.22)	135.22 (± 14.44)	136.99 (± 14.07)	135.29 (± 16.19)	133.73 (± 14.45)
Diastolic	76.87 (± 10.36)	77.57 (± 9.19)	75.72 (± 10.34)	75.97 (± 8.72)	76.88 (± 10.51)	76.20 (± 9.22)
Blood_glucose	164.63 (± 78.84)	148.84 (± 50.50)	146.02 (± 48.59)	150.81 (± 62.27)	162.97 (± 74.99)	146.05 (± 44.79)
Hba1c	7.65 (± 1.58)	7.16 (± 1.16)	7.28 (± 1.30)	7.35 (± 1.41)	7.69 (± 1.24)	7.12 (± 1.12)
Serum creatinine	1.37 (± 0.46)	1.06 (± 0.33)	1.08 (± 0.31)	1.13 (± 0.41)	1.29 (± 0.46)	1.16 (± 0.39)
Serum_cholesterol	194.03 (± 53.31)	184.37 (± 51.58)	181.45 (± 43.61)	178.25 (± 43.29)	184.32 (± 42.98)	178.80 (± 43.64)
Serum_cholesterol LDL	99.58 (± 32.36)	99.09 (± 35.50)	98.68 (± 32.98)	94.09 (± 28.24)	92.11 (± 33.25)	95.27 (± 27.59)
Serum_cholesterol HDL	46.26 (± 14.85)	52.18 (± 15.84)	49.07 (± 12.77)	50.21 (± 15.96)	46.94 (± 22.31)	49.81 (± 13.70)
Serum_triglycerides	211.39 (± 124.06)	176.37 (± 103.07)	172.66 (± 119.14)	180.36 (± 168.67)	247.88 (± 152.54)	167.35 (± 81.73)
Serum_potassium	4.61 (± 0.60)	4.56 (± 0.48)	4.55 (± 0.46)	4.53 (± 0.50)	4.62 (± 0.60)	4.52 (± 0.55)
Hemoglobin	13.64 (± 1.81)	13.48 (± 1.45)	13.74 (± 1.45)	13.39 (± 1.63)	13.93 (± 1.53)	13.33 (± 1.68)
Serum_albumin	4.38 (± 0.48)	4.51 (± 0.43)	4.53 (± 0.54)	4.44 (± 0.50)	4.57 (± 0.63)	4.48 (± 0.47)
Crp	0.51 (± 0.53)	0.61 (± 1.23)	0.51 (± 1.08)	1.01 (± 3.01)	0.48 (± 0.47)	0.51 (± 1.04)
Mean uacr	173.37 (± 379.24)	48.63 (± 124.42)	23.58 (± 33.45)	69.52 (± 180.33)	168.19 (± 364.34)	31.06 (± 66.02)

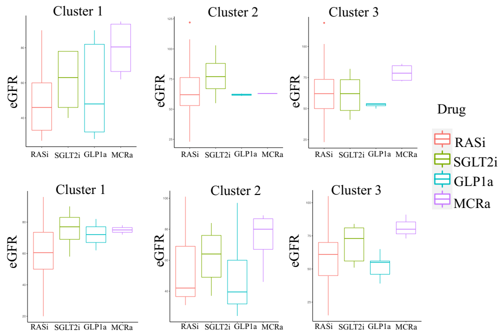

Fig. 7 Box plots of clusters of patients for the different treatments and different eGFR levels. Top: $D(t_0,t_1)$. Bottom: $D(t_1,t_2)$. Drug a_1 is RASi, a_2 is SGLT2i, a_3 is GLP1a, and a_4 is MCRa.

Acknowledgements This research activity is part of the project DC-ren that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 848011. We are grateful to the researchers of the European Centre for Living Technology (ECLT) for very helpful discussions and suggestions.

References

1. Amiri, S., Clarke, B. S., Clarke, J. L.: Clustering categorical data via ensembling dissimilarity matrices. J. Comput. Graph. Statist. **27** (1): 195–208 (2017)
2. Carlsson G., Mémoli, F.: Classifying Clustering Schemes. Foundations of Computational Mathematics **13**, 221–252 (2013)
3. Gower J.: A general coefficient of similarity and some of its properties. Biometrics **27**, 857–871 (1971)
4. Grandis, M.: Higher Category Theory. World Scientific, Singapore (2020)
5. Hummel, M., Edelmann, D., and Kopp-Schneider, A.: Clustering of samples and variables with mixed-type data. Plos One **12** (11), e0188274 (2017)
6. Park, S., Xu, H., Zhao, H.: Integrating Multidimensional Data for Clustering Analysis With Applications to Cancer Patient Data. Journal of the American Statistical Association **116** (533), 14–26 (2021)