Visualizing electronic medical records of diabetic patients using pairwise similarity for explainable structuring

Joris Falip, Sara Barraud, Frédéric Blanchard

Visualizing electronic medical records of diabetic patients using pairwise similarity for explainable structuring

Joris Falip[1][0000−0001−5115−3945], Sara Barraud[2], and Frédéric Blanchard[1]

[1] CReSTIC, Reims, France firstname.lastname@univ-reims.fr
[2] Department of Endocrinology, Diabetes and Nutrition, University Hospital of Reims, France

Abstract. As medical databases grow larger and larger, medical experts oftenlack appropriate and accessible tools to make the best of the datasetsavailable and transform data into actionable information. Manyknowledge extraction algorithms provide relevant results but failto provide explainable and transparent results. Accountability isparamount in healthcare, and hospital staff cannot rely on black boxtools when it comes to taking informed decisions. To address thissituation we propose an algorithm able to structure thousands ofelectronic medical records by similarity and typicality. Using a rank-based approach suitable for high-dimensional data, we associateeach patient's record to a very similar yet more typical record.This provides a structure suitable for data visualization, allowingfor both a high-level summary of a cohort and its representativepatients, and a detailed representation of similarities and relationsin each cluster. We applied this method to electronic medical recordsof diabetic patients, providing an easy tool for visualization andexploration of these data with the added benefit of explainability.

Keywords: Data mining · Similarity structuring · Exemplar theory

1 Introduction

Type 1 diabetes mellitus is a chronic disease characterized by insulindeficiency due to pancreatic β-cell loss and leads to hyperglycemia. Apart from the immediate life-threatening risk without insulin treatment, diabetes can have a long-term impact on the life expectancy and quality of life of these patients[2]. The main complications are retinopathy, kidney disease, neuropathy, and cardiovascular diseases. The incidence of type 1 diabetes has been increasing in France and worldwide, particularly in children, thus increasing the number of patients exposed to the risk of complications[6,7]. The main known marker of this risk is the HbA1c level, an estimation of the average blood sugar level over 2-3 months. However, the risk seems to be different between patients with the same HbA1c level, and other factors like exposure time to hyperglycemia seem to be involved.

In order to limit the complications of diabetes and its impact on life expectancy and quality of life, it is necessary to better understand the factors favoring these complications. In the Champagne-Ardenne area, for more than 15 years, information on the follow-up of diabetic patients has been recorded in the CAéDIAB database (Champagne Ardenne Réseau Diabète) at each appointment and each hospitalization. That makes it an essential database with a long-term follow-up of patients affected by type 1 diabetes. In order to analyze this database, we built a collaboration between the CReSTIC laboratory, the Endocrinology-Diabetes-Nutrition department of Reims University Hospital, and the ORNICARE (CARéDIAB) health network. This collaboration aims to explore the risk factors of diabetes complications from the CARéDIAB database, using machine learning algorithms.

Given the amount of data, medical experts must rely on unsupervised machine learning to carry out interactive exploratory data analysis. For such a task we cannot require any *a priori* assumptions but must offer a way to account for user preferences in the process. Results also need to be explainable and interpretable to maximize their utility for the hospital staff[5]. Medical data also tend to be high-dimensional data: every element of the dataset is described by a large number of features. This comes with two inherent problems: when the number of dimensions grows, data becomes sparser as the density decreases at any point in the description space, and elements tend to become equidistant and thus harder to compare using traditional distances[3].

2 Structuring around exemplars

Our approach revolves around "exemplars": typical and representative individuals that can subsume parts of the dataset. By extracting the exemplars from a dataset or database, we can provide a high-level summary of the studied population. To reach this goal, we rely on a method where each dimension is processed individually. Uni-dimensional results are then aggregated in a final step. By using this approach and relying heavily on the ranking of elements, we can provide meaningful results even with high-dimensional data as we do not use distance in the full description space. Ranks also have the added benefit of not excluding outliers.

The proposed algorithm relies on the *Degree of Representativeness*[1] (*DoR*) of the elements. The *DoR* quantifies the ability of an individual to act as an exemplar and subsume other individuals. To establish the structure, we create a directed graph where each element is connected to its exemplar: its neighbor with the highest *DoR*. In the resulting graph, each vertex represents a patient's electronic medical record, each edge represents a strong similarity to a more typical exemplar, and connected components can be summarized by their most central exemplar.

Let Ω be a set of N objects in a multidimensional space. These N objects, or elements, are characterized by D qualitative or quantitative features.

For each of the D dimensions, we compute the distance matrix. Each object then ranks every other object according to their similarity on this dimension: low distance translates to a good ranking, with the nearest element being ranked 1 and the farthest ranked $N-1$. This step is repeated on each dimension to transform the D distance matrices into D rank matrices.

Let us transform the ranks into scores. Let x be an object of Ω: for each dimension d, x assigns a relative score $Score_x^d$ to everyother object y of Ω. $Score_x^d$, relative to x, can be any arbitrary function, but in this paper it will be defined by:

$$Score_x^d(y) = max(1 + T - Rank_x^d(y), 0)$$

where $Rank_x^d(y)$ is the rank of an object y relative to x on dimension d, and each element only assigns a non-zero score to its T nearest neighbors. For each element y, we can compute the sum of all scores it received on a specific dimension:

$$Score^d(y) = \sum_{x \in \Omega} Score_x^d(y)$$

Let us define k as a numeric parameter allowing control over the total number of exemplars. To find the exemplar of a given object x, we introduce the $DoR_x(y)$ of another element y as the sum of the scores $Score^d(y)$ for every dimension d where y is inthe k nearest neighbors of$x$.

$$DoR_x(y) = \sum_{d \in D'} Score^d(y)$$

where D' is the set of dimensions on which y is in the k-nearest neighbors of x. The element chosen as an exemplar for object x is the element with the highest DoR_x.

3 Conclusion

The structure resulting from our algorithm can be visualized to help with data manipulation and exploration, so medical experts have an easier time understanding how datasets are structured and why some patients experience the same evolution of a disease (see Figure 1). Thanks to our approach where the *Degree of Representativeness* of each element is the sum of its score on various dimensions, we can provide an explanation of each edge by describing which dimensions resulted in a similarity between two elements. We compared our results to various methods used to structure datasets, using both simulated and real world data. Our results outlined better performance and a method less prone to the inherent problems of high-dimensionality usually encountered with these datasets.

From a practical point of view, this method is currently available as a recommender system used by medical experts studying diabetic patients[4] to assist them in epidemiological studies and patient's diagnostic. Future works include further development of our currently deployed prototype to easily gather user feedback on each association between elements. This feedback could be used to create an automatic weighing of the features, tailoring recommendations to the user's preferences.

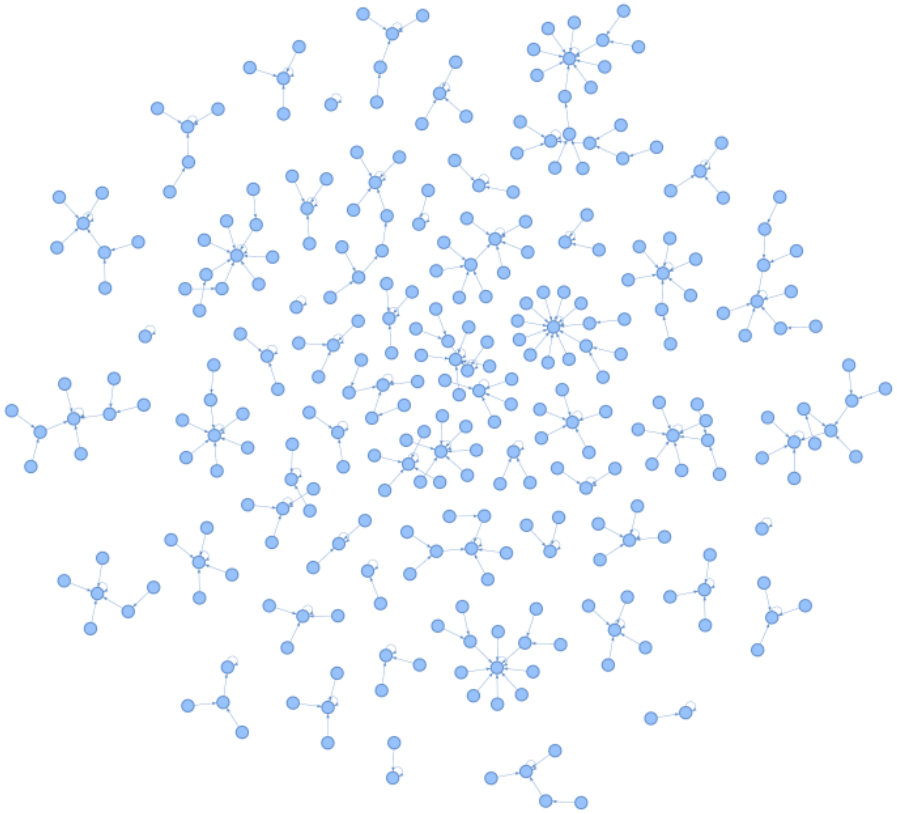

Fig. 1. Graph representing the similarities between electronic medical records of diabetic patients. Patients in the same connected components followed a similar evolution of the disease, with central ones being more typical.

References

1. Frédéric Blanchard, Amine Aït-Younes, and Michel Herbin. 2015. *Linking Data According to Their Degree of Representativeness (DoR)*. EAI Endorsed Transactionson Industrial Networks and Intelligent Systems 2, 4 (June 2015), e2
2. Linda A DiMeglio, Carmella Evans-Molina, and Richard A Oram. 2018. *Type 1 diabetes*. The Lancet 391, 10138 (June 2018), 2449–2462.
3. David L Donoho et al. 2000. *High-dimensional data analysis: The curses andblessings of dimensionality*. AMS Math Challenges Lecture 1 (2000), 32.
4. Joris Falip, Amine Aït Younes, Frédéric Blanchard, Brigitte Delemer, Alpha Diallo, and Michel Herbin. 2017. *Visual instance-based recommendation system for medical data mining*. In KES. 1747–1754.
5. Zachary C Lipton. 2016. *The mythos of model interpretability*. arXiv preprint:1606.03490.
6. David M. Maahs, Nancy A. West, Jean M. Lawrence, and Elizabeth J. Mayer-Davis. 2010. *Epidemiology of type 1 diabetes*. Endocrinology and Metabolism Clinics of North America 39, 3 (Sept. 2010), 481–497.
7. C Piffaretti, L Mandereau-Bruno, S Guilmin-Crepon, C Choleau, R Coutant, and S Fosse-Edorh. 2017. *Incidence du diabète de type 1 chez l'enfant en France en 2013-2015, à partir du système national des données de santé (SNDS). Variations régionales*. Bulletin épidémiologique hebdomadaire 27-28 (Nov. 2017).