Omega 107 (2022) 102543

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Interpreting clusters via prototype optimization☆

Emilio Carrizosa[a], Kseniia Kurishchenko[b,*], Alfredo Marín[c], Dolores Romero Morales[b]

[a] Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[b] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark
[c] Departamento de Estadística e Investigación Operativa, Universidad de Murcia, Murcia, Spain

ARTICLE INFO

Article history:
Received 12 February 2021
Accepted 8 September 2021
Available online 23 September 2021

Keywords:
Machine Learning
Interpretability
Cluster Analysis
Prototypes
Mixed-Integer Programming

ABSTRACT

In this paper, we tackle the problem of enhancing the interpretability of the results of Cluster Analysis. Our goal is to find an explanation for each cluster, such that clusters are characterized as precisely and distinctively as possible, i.e., the explanation is fulfilled by as many as possible individuals of the corresponding cluster, *true positive* cases, and by as few as possible individuals in the remaining clusters, *false positive* cases. We assume that a dissimilarity between the individuals is given, and propose distance-based explanations, namely those defined by individuals that are close to its so-called prototype. To find the set of prototypes, we address the biobjective optimization problem that maximizes the total number of true positive cases across all clusters and minimizes the total number of false positive cases, while controlling the true positive rate as well as the false positive rate in each cluster. We develop two mathematical optimization models, inspired by classic Location Analysis problems, that differ in the way individuals are allocated to prototypes. We illustrate the explanations provided by these models and their accuracy in both real-life data as well as simulated data.

1. Introduction

With the growing popularity of machine learning methods in data driven decision making, their complexity is increasing too. This may harm interpretability, a desirable property that is sought in many domains, e.g., credit scoring, medical diagnosis, and regulatory benchmarking [1–7], but also imposed in the European Union's new General Data Protection Regulation (GDPR) [8] when citizens are subject to algorithmic decision making. There have been some attempts to enhance the interpretability of Supervised Learning methods [9,10], e.g., an interpretable version of random forest [11], support vector machines [12], and deep learning [13]. This paper is devoted to the interpretability of one of the most popular Unsupervised Learning methods, namely, Cluster Analysis [14]. The need of interpretability in Cluster Analysis arises in many applications, such as security [15], internet traffic [16], finance [17], sales profiling [18], and astronomy [19].

There are two ways of enhancing interpretability in Cluster Analysis: intrinsic models and post-hoc models. Intrinsic models build simultaneously clusters and their explanations [20,21], while post-approaches are needed to interpret existing clusters, that have been built in the past, and for which we only have a label for each individual. Throughout this section, we will use a running example with clusters given, namely the real-world dataset containing 12 countries about the opinions of political science students, see Table 1. In [22], three clusters are given for this dataset, cluster 1 composed by Belgium, Egypt, France, Israel, and USA; cluster 2 with Brasil, India, and Zaire; and cluster 3 with China, Cuba, USSR, and Yugoslavia.

There are some works in the literature on post-hoc approaches. In [23], the authors assume that the individuals have been evaluated on a set of features and propose rule-based explanations. There are also ad-hoc approaches as those in, e.g., [24–26], for specific types of data. In this paper, we propose a post-hoc approach for interpreting clusters via means of prototypes.

Our starting point is the predefined clusters in $\mathcal{C}$, which have been obtained applying a clustering procedure to the set of individuals $\mathcal{N}$ [27–33]. We propose a methodology to improve the interpretability of the results of Cluster Analysis, by giving an explanation to each cluster $c \in \mathcal{C}$ that characterizes as precisely and distinctively as possible c. In other words, the explanation is to be

☆ Area: Data-Driven Analytics. This manuscript was processed by Associate Editor Joe Zhu.

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), kk.eco@cbs.dk (K. Kurishchenko), amarin@um.es (A. Marín), drm.eco@cbs.dk (D. Romero Morales).

https://doi.org/10.1016/j.omega.2021.102543

Table 1
Dissimilarities on opinions of political science students between the 12 countries in our running example, [22].

Country	Dissimilarities to other countries										
	Belgium	Brasil	China	Cuba	Egypt	France	India	Israel	USA	USSR	Yugoslavia
Brasil	5.58										
China	7.00	6.50									
Cuba	7.08	7.00	3.83								
Egypt	4.83	5.08	8.17	5.83							
France	2.17	5.75	6.67	6.92	4.92						
India	6.42	5.00	5.58	6.00	4.67	6.42					
Israel	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USSR	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
Yugoslavia	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
Zaire	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

fulfilled by as many as possible individuals of c (and these will be referred to as *true positive* cases) and by as few as possible individuals in the remaining clusters (which will be referred to as *false positive* cases).

Our explanations are distance-based, as in clustering procedures attempting to partition the set of individuals such that individuals that are close to each other are allocated to the same cluster, whereas individuals that are far from each other are expected to be in different clusters. It is then natural to *explain* cluster c following a distance-based explanation such as

c is the set of individuals of $\mathcal{N}$ that are close to a given individual i.

To define distance-based explanations, we assume we are given a dissimilarity δ to measure the closeness between individuals [34]. The dissimilarity between the 12 countries in our running example is given in Table 1. Note that, in general, δ does not need to be the dissimilarity used to construct the clusters in $\mathcal{C}$. Actually, that dissimilarity may not be available to us.

How well this explains cluster c depends on the choice of individual i to which we will refer as the prototype of cluster c [35,36], in other words, the "face" chosen for the cluster. Our aim is to select the set of prototypes that maximizes the total number of true positive cases across all clusters and minimizes the total number of false positive cases while controlling the true positive rate as well as the false positive rate in each cluster. With the methodology proposed in this paper, the chosen prototypes for our example are: France for cluster 1, Brasil for cluster 2, and Yugoslavia for cluster 3. For cluster 1, all 5 countries are true positive cases, while none of the 7 countries in the other two clusters are false positive cases, yielding to the ideal quality of the explanation, namely 100% true positive rate and 0% false positive rate. The same holds for the other two clusters.

In general, one cannot expect to find perfect explanations. In Fig. 1, we can see that by trying to improve the number of true positive cases of an explanation we may harm the number of false positive cases. There we have two clusters, cluster 1 with 5 individuals represented by a red star and cluster 2 with 4 individuals represented by a blue star. If we look at the explanation in Fig. 1a for cluster 1, the circle in red containing 4 of the individuals from cluster 1 and none from cluster 2, we see that there are 4 true positive cases (or, equivalently, an 80% true positive rate) and 0 false positive cases (or, equivalently, a 0% false positive rate), while for the alternative explanation for cluster 1 in Fig. 1b, the number of true positive cases has increased to 5 (achieving a 100% true positive rate) but the number of false positive cases has gone up to 1 (25% false positive rate).

To find the set of prototypes, we propose two mathematical optimization models, the covering and the partitioning ones, inspired by classic Location Analysis problems, namely the covering [37] and the p-median problems [38,39]. In the covering model, a cluster is explained as the individuals whose distance to its prototype is below a threshold value, i.e., the explanation of cluster c can be visualized as the ball in the distance δ centered at its prototype and radius equal to the corresponding threshold value. Instead, in the set-partitioning model, cluster c is explained as the individuals that are the closest to the prototype of c than to the prototypes of the other clusters. In this case, the explanations can be visualized as Voronoi diagrams. For both models, we provide a Mixed Integer Linear Programming (MILP) formulation, where in the covering one, in addition to the prototypes, we need to decide the size of the radii.

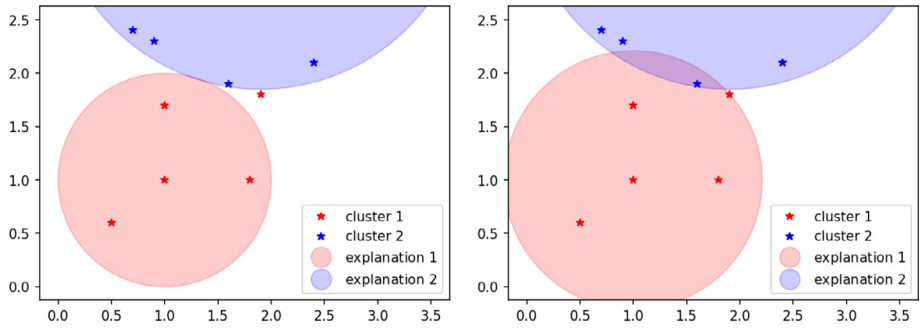
The remainder of the paper is organized as follows. Section 2 presents the covering model, while Section 3 the partitioning model. Section 4 provides numerical results for real-life data as well as simulated data. Section 5 summarizes the paper and proposes future lines of research.

2. The covering model

In this model, given a cluster c, a prototype i, an individual will be considered covered by cluster c if it is close enough to i. By close enough we mean that their dissimilarity is below a threshold value r_c, which is the coverage radius. Our aim is thus to find the prototypes and the cluster radii. Observe that, with this approach, an individual could be covered by more than one cluster if some of the radii are large, while some individuals may not be covered by any cluster when the radii are small. We obtain an MILP formulation for this problem, which is separable on the clusters. We show how the radii can only take on a discrete amount of values, and give an alternative Integer Programming (IP) formulation for a fixed radius. We focus on the most interpretable case in which only one prototype per cluster is to be selected. The extension to more than one prototype is straightforward.

Let us introduce the problem more formally. We are given a clustering $\mathcal{C}$ obtained from splitting the individuals in $\mathcal{N}$, $\mathcal{N} = \bigcup_{c\in\mathcal{C}}\mathcal{N}_c$. The prototype of cluster c is chosen from set $\mathcal{I}_c \subseteq \mathcal{N}_c$, with $\mathcal{I} = \bigcup_{c\in\mathcal{C}}\mathcal{I}_c$. We are also given the dissimilarity between prototype i and individual n, δ_{in}, for every $i \in \mathcal{I}$ and $n \in \mathcal{N}$. This dissimilarity does not need to be the one that was used to construct the clusters. As pointed out in the introduction, we may have been given only clusters, and neither the method nor the dissimilarity used to build them.

Let r_c be the radius of the explanation chosen for cluster c. For $i \in \mathcal{I}_c$, let π_{in} be the binary decision variable which takes on the value 1 if $n \in \mathcal{N}$ lies in the ball of radius r_c centered at prototype $i \in \mathcal{I}$, and 0 otherwise. Moreover, let z_i be the binary decision variable which takes on the value 1 if i is chosen as prototype and 0 otherwise. Throughout the paper, we use bold typesetting to denote the vectors, e.g., $\mathbf{r} = (r_c)_{c\in\mathcal{C}}$.



(a) An explanation for cluster 1 (b) An alternative explanation for cluster 1

Fig. 1. Illustration of the trade-off between true positive and false positive cases.

With these variables, the number of *true positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i$ and the True Positive Rate (TPR_c) is

$$\text{TPR}_c = \frac{\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i}{|\mathcal{N}_c|}, \tag{1}$$

while the number of *false positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i$ and the False Positive Rate (FPR_c) is

$$\text{FPR}_c = \frac{\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i}{|\mathcal{N}\setminus\mathcal{N}_c|}. \tag{2}$$

The covering model reads as follows:

$$\max_{\mathbf{z},\boldsymbol{\pi},\mathbf{r}} \quad \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i - \theta\sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i \tag{3}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}_c} z_i = 1,\ \forall c\in\mathcal{C} \tag{4}$$

$$r_c \geq \delta_{in}\pi_{in},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}_c, \forall c\in\mathcal{C} \tag{5}$$

$$r_c \leq \delta_{in} + (r_c^{\max} - \delta_{in})\pi_{in},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}\setminus\mathcal{N}_c, \forall c\in\mathcal{C} \tag{6}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i \geq \lceil\lambda_c|\mathcal{N}_c|\rceil,\ \forall c\in\mathcal{C} \tag{7}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i \leq \lfloor\mu_c|\mathcal{N}\setminus\mathcal{N}_c|\rfloor,\ \forall c\in\mathcal{C} \tag{8}$$

$$r_c^{\min} \leq r_c \leq r_c^{\max},\ \forall c\in\mathcal{C} \tag{9}$$

$$z_i\in\{0,1\},\ \forall i\in\mathcal{I}_c, \forall c\in\mathcal{C} \tag{10}$$

$$\pi_{in}\in\{0,1\},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C}. \tag{11}$$

The objective function is equal to the total number of true positive cases across all clusters minus the total number of false positive cases weighted by the trade-off parameter $\theta\geq 0$. Constraints (4) ensure that one single prototype is chosen for each cluster. Constraints (5) and (6) ensure that the decision variables π_{in} are well defined. Note that because of the shape of the objective function, for $n\in\mathcal{N}_c$, we only need to ensure that if $r_c < \delta_{in}$ then $\pi_{in}=0$, which is done by constraint (5). For $n\in\mathcal{N}\setminus\mathcal{N}_c$, we only need to ensure that if $r_c > \delta_{in}$ then $\pi_{in}=1$, which is done by constraints (6). Note that if $r_c=\delta_{in}$ then $\pi_{in}=1$ for individuals inside the cluster c and $\pi_{in}=0$ for individuals outside the cluster c. It is easy to see that constraints (7) control the true positive rate in cluster c, TPR_c, via the parameter $\lambda_c\in[0,1]$. Similarly, constraints (8) control the false positive rate in cluster c, FPR_c, via the parameter $\mu_c\in[0,1]$. Finally, constraints (9)–(11) define the nature of the decision variables. The radius of cluster c is bounded from below and above by $r_c^{\min}$ and $r_c^{\max}$, respectively. Straightforward values for these parameters are $r_c^{\min}=\min_{(i,n)\in\mathcal{I}_c\times\mathcal{N}_c, i\neq n}\delta_{in}$ and $r_c^{\max}=\max_{(i,n)\in\mathcal{I}_c\times\mathcal{N}_c}\delta_{in}$.

Note that the objective function contains the total number of true and false positive cases across all clusters, while constraints (7)–(8) allow us to control these two criteria in each cluster. These constraints can be useful when we want to prioritize how well we explain certain clusters, or when the clusters are of very different size and we want to ensure a good performance independently of their size, as we do in the numerical section for the real-world dataset.

In formulation (3)–(11), we have the product of two decision variables, i.e., π_{in} and z_i, which makes the problem bi-linear. We can obtain an equivalent MILP formulation, by applying the Fortet transformation [40]. Let us introduce the new decision variable $y_{in}=\pi_{in}z_i$ and the following constraints to ensure y_{in} is well-defined:

$$y_{in}\leq\pi_{in},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C} \tag{12}$$

$$y_{in}\leq z_i,\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C} \tag{13}$$

$$y_{in}\geq\pi_{in}+z_i-1,\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C} \tag{14}$$

$$y_{in}\in\{0,1\},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C}. \tag{15}$$

The covering model (3)–(15) has been formulated as an MILP with $2|\mathcal{I}|\times|\mathcal{N}|+|\mathcal{I}|$ binary and $|\mathcal{C}|$ continuous decision variables, and $4|\mathcal{I}|\times|\mathcal{N}|+4|\mathcal{C}|$ linear constraints. Note that this MILP formulation is separable on the clusters. Indeed, the objective function consists of a summation across the clusters of the number of true positive cases minus the number of false positive cases weighted by θ. Similarly, the constraints relevant to c only involve decision variables relating to c.

We have modeled the radius of cluster c, r_c, as a continuous variable. However, it is easy to show that we only need to consider a discrete amount of values, namely, $r_c\in\{\delta_{in},\ \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}_c\}$. Suppose that we solve the covering model for one of these values. Since the radius is fixed, the values of π_{in} are known and can

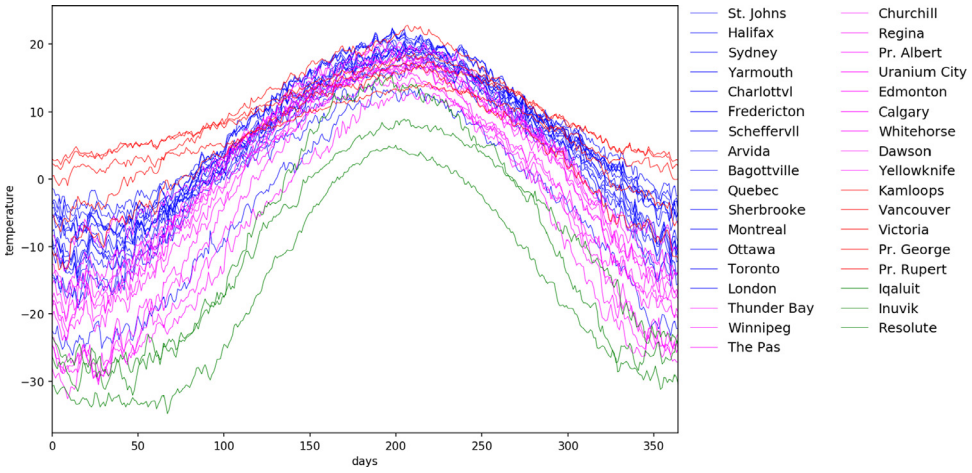


Fig. 2. The Canadian weather data grouped into four clusters by climate's type: Atlantic - blue, Continental - pink, Pacific - red, Arctic - green. Days are along the horizontal axis, temperatures are along the vertical axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be calculated in a preprocessing step, as well as the true positive cases and false positive cases associated with i if i is chosen as a prototype.

Let us denote by π_{in}^r the value of π_{in} when the radius of cluster c, r_c, is fixed to r. Let us define

$$\phi_{ic}^r = \sum_{n \in \mathcal{N}_c} \pi_{in}^r,$$

$$\psi_{ic}^r = \sum_{n \in \mathcal{N} \setminus \mathcal{N}_c} \pi_{in}^r.$$

With this, the covering model for cluster c and radius $r_c = r$ can be formulated as follows:

$$\max_{\mathbf{z}} \quad \sum_{i \in \mathcal{I}_c} \phi_{ic}^r z_i - \theta \sum_{i \in \mathcal{I}_c} \psi_{ic}^r z_i \tag{16}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_c} z_i = 1 \tag{17}$$

$$\sum_{i \in \mathcal{I}_c} \phi_{ic}^r z_i \geq \lceil \lambda_c |\mathcal{N}_c| \rceil, \tag{18}$$

$$\sum_{i \in \mathcal{I}_c} \psi_{ic}^r z_i \leq \lfloor \mu_c |\mathcal{N} \setminus \mathcal{N}_c| \rfloor, \tag{19}$$

$$z_i \in \{0, 1\}, \ \forall i \in \mathcal{I}_c. \tag{20}$$

Note that the set of candidates to prototype for cluster c, $\mathcal{I}_c$, can be reduced to $\mathcal{I}_c' \subset \mathcal{I}_c$. Some candidates can be removed because $\phi_{ic}^r < \lceil \lambda_c |\mathcal{N}_c| \rceil$ and others because $\psi_{ic}^r > \lfloor \mu_c |\mathcal{N} \setminus \mathcal{N}_c| \rfloor$. After reducing the set of candidates from $\mathcal{I}_c$ to $\mathcal{I}_c'$, we can eliminate constraints (18) and (19), and the problem is equivalent to choosing the prototype from $\mathcal{I}_c'$ with the largest $\phi_{ic}^r - \theta \psi_{ic}^r$.

To tackle large instances of the problem, i.e., with many individuals, we can combine our covering model with a sampling procedure from the set of individuals and/or the set of candidates to prototype. Indeed, we can sample from the set of candidates to prototype for cluster c, yielding $\tilde{\mathcal{I}}_c \subset \mathcal{I}_c$, for all c, and/or sample from the set of individuals from cluster c, yielding $\tilde{\mathcal{N}}_c \subset \mathcal{N}_c$, and solve the reduced covering model. Let z_i^{R} and r_c^{R}, $i \in \tilde{\mathcal{I}}_c$ and $c \in \mathcal{C}$, be the chosen prototypes and the chosen radii of the reduced problem if this is feasible. We can use this partial solution to find a feasible solution to the original problem, $(\mathbf{z}^{\mathrm{O}}, \boldsymbol{\pi}^{\mathrm{O}}, \mathbf{r}^{\mathrm{O}})$ with $\mathbf{z}^{\mathrm{O}} = \mathbf{z}^{\mathrm{R}}$ and $\mathbf{r}^{\mathrm{O}} = \mathbf{r}^{\mathrm{R}}$, satisfying constraints (7), imposing a lower bound on TPR_c, and constraints (8), imposing an upper bound on FPR_c. Needless to say that this approach may not yield a feasible solution to the original problem, and we may need to sample more or make the values of λ_c and μ_c less restrictive.

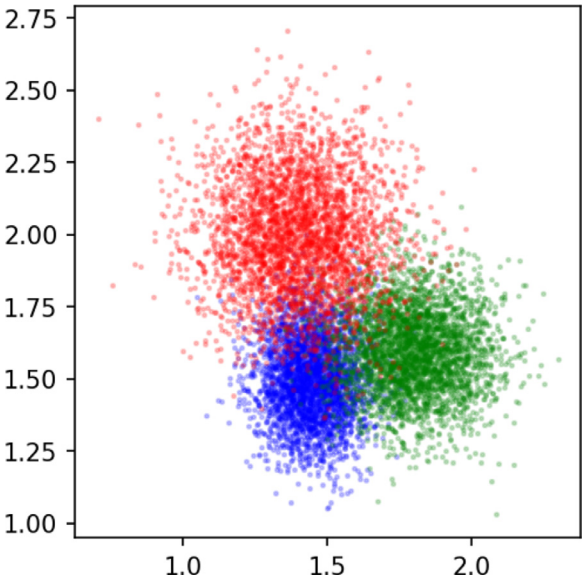
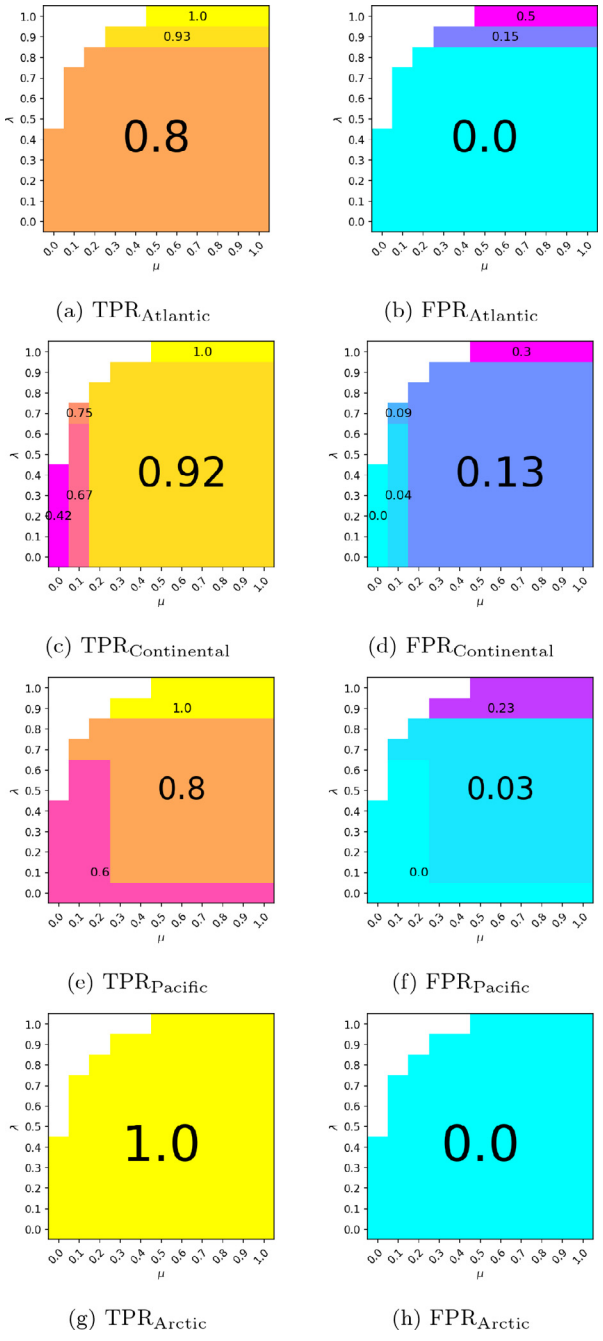


Fig. 3. Simulated data in $\mathbb{R}^2$ with three clusters.

3. The partitioning model

An alternative way of explaining clusters by means of prototypes is the partitioning model. In this case, each individual is assigned to exactly one prototype, namely the closest one. To do this, in addition to the z_i variables defined as before, we also need the binary variables ρ_{in} that allocate individuals to prototypes. Let ρ_{in} take on the value 1 if prototype i is the closest one to individual n from the chosen ones, and 0 otherwise. With these variables, the number of *true positive* cases in cluster c is equal to $\sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in}$ and

$$\mathrm{TPR}_c = \frac{\sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in}}{|\mathcal{N}_c|}, \tag{21}$$



(a) $\text{TPR}_{\text{Atlantic}}$ (b) $\text{FPR}_{\text{Atlantic}}$

(c) $\text{TPR}_{\text{Continental}}$ (d) $\text{FPR}_{\text{Continental}}$

(e) $\text{TPR}_{\text{Pacific}}$ (f) $\text{FPR}_{\text{Pacific}}$

(g) $\text{TPR}_{\text{Arctic}}$ (h) $\text{FPR}_{\text{Arctic}}$

Fig. 4. For each cluster of the Canadian weather data, the true positive ratio and false positive ratio given by the covering model when λ and μ vary on a grid in $[0,1]\times[0,1]$.

while the number of *false positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in}$ and

$$\text{FPR}_c = \frac{\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in}}{|\mathcal{N}\setminus\mathcal{N}_c|}. \tag{22}$$

The partitioning model reads as follows:

$$\max_{\mathbf{z},\boldsymbol{\rho}} \quad \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\rho_{in} - \theta\sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in} \tag{23}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}_c} z_i = 1,\ \forall c\in\mathcal{C} \tag{24}$$

$$\sum_{j\in\mathcal{I}_c:\,\delta_{jn}\le\delta_{in}} z_j + \sum_{j\in\mathcal{I}:\,\delta_{jn}>\delta_{in}} \rho_{jn} \le 1,\ \forall(i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C} \tag{25}$$

$$\rho_{in} \le z_i,\ \forall(i,n)\in\mathcal{I}\times\mathcal{N} \tag{26}$$

$$\sum_{i\in\mathcal{I}}\rho_{in} = 1,\ \forall n\in\mathcal{N} \tag{27}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\rho_{in} \ge \lceil\lambda_c|\mathcal{N}_c|\rceil,\ \forall c\in\mathcal{C} \tag{28}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in} \le \lfloor\mu_c|\mathcal{N}\setminus\mathcal{N}_c|\rfloor,\ \forall c\in\mathcal{C} \tag{29}$$

$$z_i\in\{0,1\},\ \forall i\in\mathcal{I} \tag{30}$$

$$\rho_{in}\in\{0,1\},\ \forall(i,n)\in\mathcal{I}\times\mathcal{N}. \tag{31}$$

The objective function (23) is as in the covering model, as well as constraints (24) ensuring that we choose exactly one prototype for cluster c and constraints (28)-(29) controlling TPR_c and FPR_c for all $c\in\mathcal{C}$. Constraints (25) are the closest assignment constraints and reinforce [41] using the fact that, for each cluster, only one prototype is chosen. These constraints make sure that if individual n is assigned to a prototype, then there cannot be another prototype closer to n. Constraints (26) ensure that individuals are assigned to prototypes that have been selected. Constraints (27) impose that the model assigns each individual to a single prototype. Constraints (30)–(31) define the nature of the decision variables. Note that the integrality constraint on variable ρ_{in} can be relaxed to $\rho_{in}\ge 0$ without loss of optimality, while in the objective function it is enough to maximize $\sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\rho_{in}$ thanks to constraints (27). The partitioning model (23)-(31) has been written as an MILP problem with $|\mathcal{I}|\times|\mathcal{N}|+|\mathcal{I}|$ binary decision variables and $2|\mathcal{I}|\times|\mathcal{N}|+3|\mathcal{C}|+|\mathcal{N}|$ linear constraints.

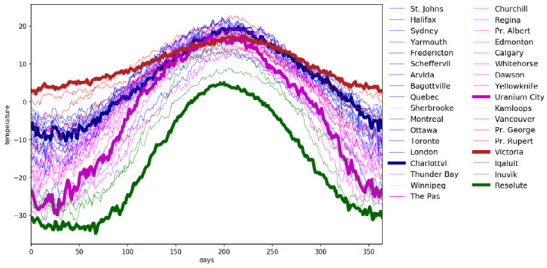
In the model above we have chosen one prototype per cluster. If we were to choose more than one, we will obviously need to change the right-hand side of constraints (24), as well as replace (25) by the original [41] constraints

$$z_i + \sum_{j\in\mathcal{I}:\delta_{in}<\delta_{jn}} \rho_{jn} \le 1,\ \forall(i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C}.$$

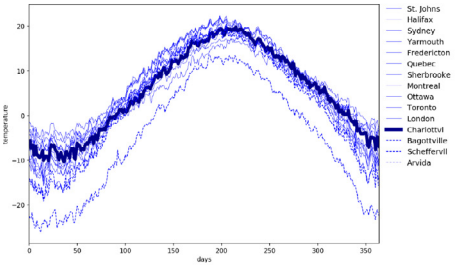
Note that there is a clear difference between the partitioning model (23)-(31) and the covering model introduced in the previous section. To define the explanations in the partitioning model, we need to know the prototypes for all clusters, while with the covering model, due to its separability on the clusters, we can obtain explanations for one single cluster without knowing prototypes from other clusters. Nevertheless, to tackle large instances of the problem with many individuals, we can use a similar approach as in Section 2, namely, we can reduce the size of the model that finds the prototypes by sampling in the set of individuals and/or the set of candidates to prototype.

4. Numerical results

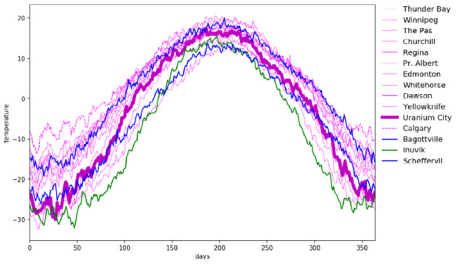
In this section, we illustrate the quality of the cluster explanations provided by the covering and the partitioning models using both real-life data and simulated data. We measure the goodness of cluster explanations by the true positive ratio TPR_c and the false positive ratio FPR_c in each of the clusters, defined in (1) and (2) for the covering problem and in (21) and (22) for



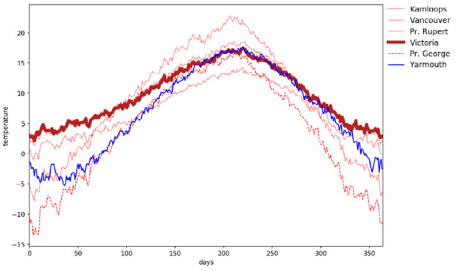
(a) The prototypes of the covering model for $\lambda = 0.80$ and $\mu = 0.20$



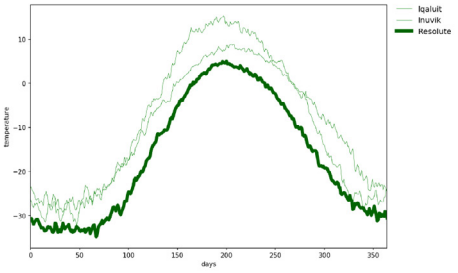
(b) $\text{TPR}_{\text{Atlantic}} = 0.80$, $\text{FPR}_{\text{Atlantic}} = 0.00$



(c) $\text{TPR}_{\text{Continental}} = 0.92$, $\text{FPR}_{\text{Continental}} = 0.13$



(d) $\text{TPR}_{\text{Pacific}} = 0.80$, $\text{FPR}_{\text{Pacific}} = 0.03$



(e) $\text{TPR}_{\text{Artic}} = 1.00$, $\text{FPR}_{\text{Artic}} = 0.00$

Fig. 5. The chosen prototypes for the Canadian weather dataset highlighted in boldface, with $\lambda = 0.80$ and $\mu = 0.20$, for the covering model. The lines of the same color as the cluster denote true positive cases; the lines of color different from the one of the cluster denote false positive cases; the dashed lines of the same color as the cluster denote false negative cases.

the partitioning problem. The explanations are obtained assuming that $\lambda = \lambda_1 = \ldots = \lambda_{|\mathcal{C}|}$ and $\mu = \mu_1 = \ldots = \mu_{|\mathcal{C}|}$. This means that throughout this section, and with loss of generality, we impose the same requirements on TPR_c to all clusters, as well as on FPR_c.

We have set the parameter in the objective function of the covering model, θ, which weighs between the total number of true positive cases and false positive ones, equal to 1. This parameter does not play a role in the partitioning model as pointed out in Section 3, where we maximize the total number of true positive cases subject to the performance constraints on TPR_c and FPR_c. To illustrate the tradeoff between TPR_c and FPR_c, we vary the parameters λ and μ on a grid in $[0, 1] \times [0, 1]$.

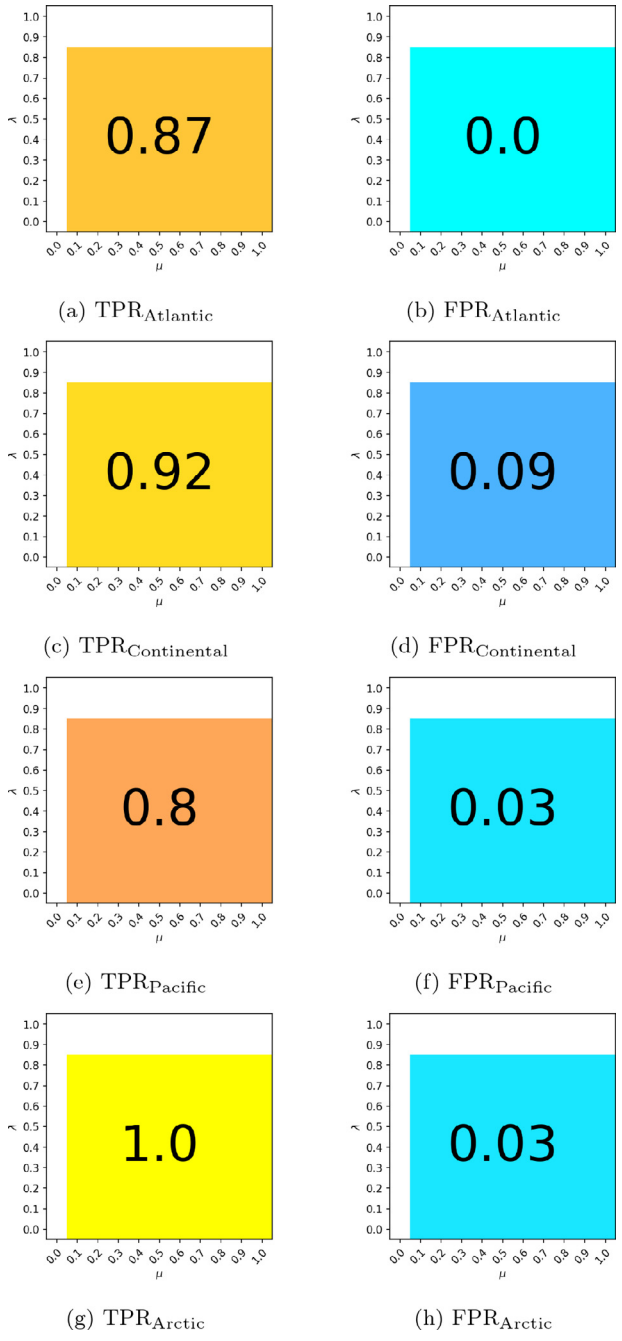
As real-life data, we use functional data relating to Canadian weather data, see Fig. 2 and Section 4.1, publicly available in the *R* package *fda* [42].

With this data we illustrate that our approach can generate good explanations, i.e., with high TPR_c and with low FPR_c, and that for some of the clusters we even obtain perfect explanations, i.e., with $\text{TPR}_c = 1$ and $\text{FPR}_c = 0$. Our grid results illustrate how by increasing the requirements on TPR_c through the parameter λ, we have to compromise the FPR_c of some clusters. In terms of simulated data, we use synthetic clusters in $\mathbb{R}^2$, see Fig. 3 and Section 4.2, and illustrate how our approach achieves good explanations in terms of TPR_c and FPR_c, even for large number of individuals $|\mathcal{N}|$.

To solve the mathematical optimization models arising we use *Gurobi* [43] with *Python* [44] on a PC Intel®Core TM i7-8665U, 16GB of RAM. We have imposed a time limit of 300 seconds to each optimization model. Within this time limit, in our numerical results below, we have been able to prove optimality or to show that the problem is infeasible.

4.1. Results for real-life data

The Canadian weather data contains 365 days of temperature observations for $|\mathcal{N}| = 35$ cities grouped into $|\mathcal{C}| = 4$ types of climates: Atlantic ($|\mathcal{N}_{\text{Atlantic}}| = 15$), Continental ($|\mathcal{N}_{\text{Continental}}| = 12$), Pacific ($|\mathcal{N}_{\text{Pacific}}| = 5$), and Arctic ($|\mathcal{N}_{\text{Artic}}| = 3$). The data are depicted in Fig. 2, where the clusters are identified by a color, namely, blue for Atlantic, pink for Continental, red for Pacific, and green for Arctic. To build the dissimilarity measure, we use a vectorial representation of each observation with the 365 daily temperatures. We measure the dissimilarity between n and i as the Euclidean distance between the corresponding vectors of



(a) $\text{TPR}_{\text{Atlantic}}$ (b) $\text{FPR}_{\text{Atlantic}}$

(c) $\text{TPR}_{\text{Continental}}$ (d) $\text{FPR}_{\text{Continental}}$

(e) $\text{TPR}_{\text{Pacific}}$ (f) $\text{FPR}_{\text{Pacific}}$

(g) $\text{TPR}_{\text{Arctic}}$ (h) $\text{FPR}_{\text{Arctic}}$

Fig. 6. For each cluster of the Canadian weather data, the true positive ratio and false positive ratio given by the partitioning model when λ and μ vary on a grid in $[0,1]\times[0,1]$.

temperatures. In both the covering and the partitioning models, we consider $\mathcal{I}=\mathcal{N}$, i.e., all individuals are candidates to prototype.

To illustrate the tradeoff between TPR_c and FPR_c for each cluster, we vary λ and μ on a grid in $[0,1]\times[0,1]$, namely, $\lambda,\mu\in\{0.0,0.1,0.2,\ldots,1.0\}$. Recall that we impose the same requirements on TPR_c as well as on FPR_c to all clusters independently of their size, avoiding thus that our approach is significantly biased towards those clusters with most individuals. The results for the covering model can be found in Fig. 4, where we report the TPR_c and the FPR_c for each cluster, separately. We use a white background to denote a combination of (λ,μ) for which the corresponding model is infeasible, i.e., no explanation can be found ensuring a TPR_c of at least λ and a FPR_c of at most μ, for each of the clusters. In general, the covering model finds good explanations, i.e., explanations that have an attractive tradeoff between TPR_c and FPR_c for all the clusters. This is the case for $(\lambda,\mu)=(0.80,0.20)$, for which $\text{TPR}_{\text{Atlantic}}=0.80$, $\text{TPR}_{\text{Continental}}=0.92$, $\text{TPR}_{\text{Pacific}}=0.80$ and $\text{TPR}_{\text{Artic}}=1.00$, while $\text{FPR}_{\text{Atlantic}}=0.00$, $\text{FPR}_{\text{Continental}}=0.13$, $\text{FPR}_{\text{Pacific}}=0.03$ and $\text{FPR}_{\text{Artic}}=0.00$.

The explanations of the covering model for $(\lambda,\mu)=(0.80,0.20)$ are depicted in Fig. 5.

In Fig. 5a we highlight in boldface the selected prototypes for each of the clusters. Figs. 5b-5e zoom in on each of the prototypes and the individuals explained by them (true positive and false positive), as well as the ones that should have been explained but were not (false negative). To visualize this, we use lines of the same color as the prototype to denote true positive cases; the lines with a color different from the one of the prototype denote false positive cases; while the dashed lines of the same color as the prototype denote false negative cases. For instance, in Fig. 5c, we can see that the prototype of the Continental climate cluster is Uranium City (in boldface pink), Dawson is a true positive (pink line), Inuvik is a false positive (green line), while Calgary is a false negative (dashed line in pink). We can see that the covering model can find more than one explanation for an individual, e.g., Inuvik is explained by the prototypes from the Continental and the Arctic clusters, or not explained at all, e.g., Calgary.

To end with the covering model we briefly discuss the range of values of TPR_c and FPR_c in Fig. 4. By definition, the higher the value of λ, i.e., the stricter we are on the minimum requirement on TPR_c for all clusters, the worse the FPR_c. For instance, for $\mu=0.10$, $\text{FPR}_{\text{Continental}}$ worsens from 0.04 to 0.09 when increasing λ. Similarly, the lower the value of μ, i.e., the stricter we are on the maximum requirement on FPR_c for all clusters, the worse the TPR_c. For instance, for $\lambda=0.70$, $\text{TPR}_{\text{Continental}}$ worsens from 0.92 to 0.75 when decreasing μ.

We now briefly discuss the results of the partitioning model for the Canadian weather data in Fig. 6. Note that in this case, the partitioning model gives for each cluster the same TPR_c and the same FPR_c for all combinations of (λ,μ) in the chosen grid for which there is a feasible solution, i.e., for $\lambda\le 0.80$ and $\mu\ge 0.10$. More detailed information on this solution can be found in Fig. 7. There we can see that, as expected, the partitioning model gives a unique explanation for each individual.

4.2. Simulated data

In this section we consider simulated data in $\mathbb{R}^2$. The simulated data consist of three clusters, see Fig. 3 where cluster 1 is depicted in blue, cluster 2 in green, and cluster 3 in red. The coordinates of the individuals in cluster c are randomly drawn from a multivariate normal distribution, $\text{N}(\boldsymbol{\beta}^c,\Sigma^c)$, with

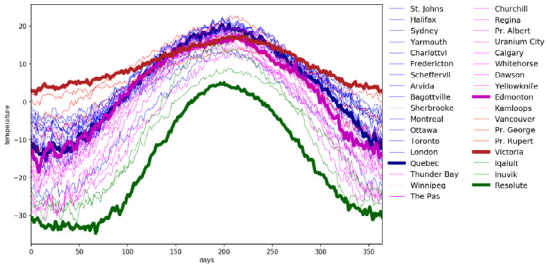
$$\boldsymbol{\beta}^1=(1.45,1.5)\qquad \boldsymbol{\beta}^2=(1.8,1.6)\qquad \boldsymbol{\beta}^3=(1.4,2.0)$$

$$\Sigma^1=\begin{pmatrix}0.01 & 0.00\\ 0.00 & 0.02\end{pmatrix}\quad \Sigma^2=\begin{pmatrix}0.02 & 0.00\\ 0.00 & 0.02\end{pmatrix}$$

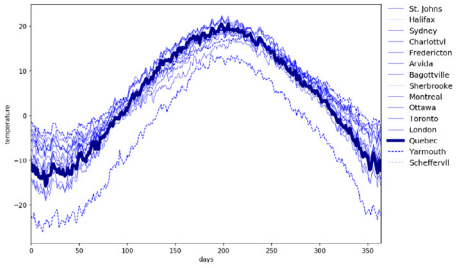
$$\Sigma^3=\begin{pmatrix}0.03 & 0.00\\ 0.00 & 0.04\end{pmatrix}.$$

We split the individuals in $\mathcal{N}$ roughly equally across the three clusters.

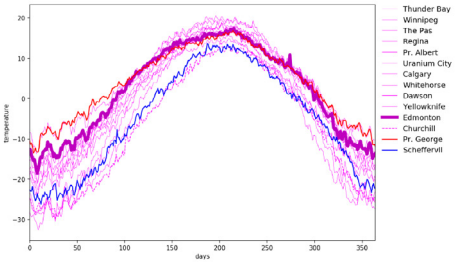
The goal of this experiment is to show that our methodology is scalable, i.e., it can handle datasets with large number of individuals and it can obtain good explanations in terms of TPR_c and FPR_c for all the clusters with both the covering and the partitioning models. For this we consider instances



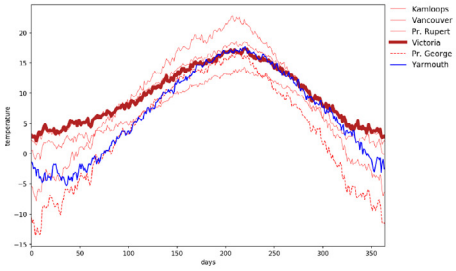
(a) The prototypes of the partitioning model for $\lambda = 0.80$ and $\mu = 0.10$



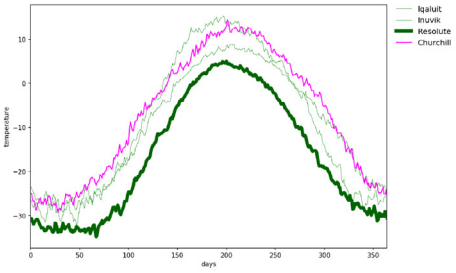
(b) $\text{TPR}_{\text{Atlantic}} = 0.87$, $\text{FPR}_{\text{Atlantic}} = 0.00$



(c) $\text{TPR}_{\text{Continental}} = 0.92$, $\text{FPR}_{\text{Continental}} = 0.09$



(d) $\text{TPR}_{\text{Pacific}} = 0.80$, $\text{FPR}_{\text{Pacific}} = 0.03$



(e) $\text{TPR}_{\text{Artic}} = 1.00$, $\text{FPR}_{\text{Artic}} = 0.03$

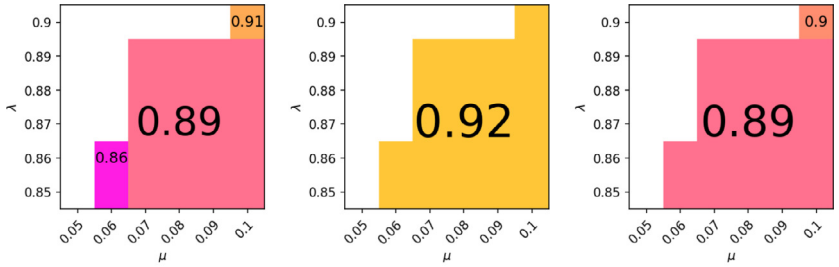
Fig. 7. The chosen prototypes for the Canadian weather dataset highlighted in boldface, with $\lambda = 0.80$ and $\mu = 0.10$, for the partitioning model. The lines of the same color as the cluster denote true positive cases; the lines of color different from the one of the cluster denote false positive cases; the dashed lines of the same color as the cluster denote false negative cases.

with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$, and we vary λ and μ on a grid in $[0, 1] \times [0, 1]$, namely, $\lambda \in \{0.85, 0.86, 0.87, 0.88, 0.89, 0.90\}$ and $\mu \in \{0.05, 0.06, 0.07, 0.08, 0.09, 0.10\}$.

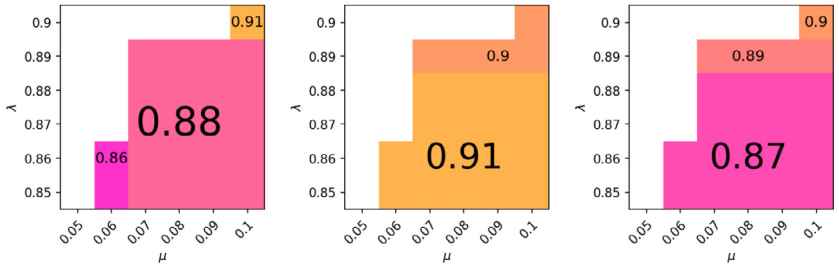
To obtain the explanations, we apply the reduction technique described in Sections 2 and 3 for the covering and the partitioning models, respectively. This consists of three steps, namely, (i) defining the data for the reduced model, (ii) finding the explanations with this new model, and (iii) evaluating the quality of the explanations in the original data. When performing (i), we select $\tilde{\mathcal{N}}_c \subset \mathcal{N}_c$ using hierarchical clustering with the Euclidean distance as the dissimilarity between the individuals in $\mathcal{N}_c$. We then choose the threshold that yields $|\tilde{\mathcal{N}}_c|$ groups of individuals. From each of these groups, we choose a representative randomly, which becomes an individual of $\tilde{\mathcal{N}}_c$. The selected individuals, with weights $\tilde{w}_n$ equal to the size of their group, across the three clusters compose $\tilde{\mathcal{N}}$. We apply a similar approach to select the individuals in $\tilde{\mathcal{I}}_c \subset \mathcal{I}_c$, for each c, by using as starting point $\tilde{\mathcal{I}}_c$ and then partition it into $|\tilde{\mathcal{I}}_c|$ groups, and select a representative randomly that becomes a member of $\tilde{\mathcal{I}}_c$. In (ii), we solve the covering and the partitioning models with individuals in $\tilde{\mathcal{N}}_c$ weighted by $\tilde{w}_n$ and candidates to prototype in $\tilde{\mathcal{I}}_c$. Third, for the obtained explanations, we calculate TPR_c and FPR_c on the original dataset $\mathcal{N}$, with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$. In the numerical results below, we take $|\tilde{\mathcal{N}}_c| = 125$ and $|\tilde{\mathcal{I}}_c| = 25$, $c = 1, 2, 3$.

We now discuss the results for the covering model, see Figs. 8 and 9. We can see that the explanations obtained with the reduced problem show a good performance on the original dataset even when the number of individuals is very large, namely $|\mathcal{N}| = 10^6$. To illustrate this, let us start with $(\lambda, \mu) = (0.90, 0.10)$. In terms of true positive cases, for $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$, we have TPR_c equal to 0.91, 0.90, 0.90, for $c = 1, 2, 3$. In terms of false positive cases, for $|\mathcal{N}| = 10^4$, we have FPR_c equal to 0.08, 0.05, 0.05, for $c = 1, 2, 3$, while for $|\mathcal{N}| = 10^5$ and 10^6, FPR_1 worsens to 0.09. This means that with the optimal solution of the reduced problem, we have been able to find explanations to the clusters that satisfy constraints (7) for $\lambda = 0.90$ and (8) for $\mu = 0.10$. For other combinations of λ and μ, the quality of the explanations provided by the reduced problem is also good, with possible minor violations of constraints (7) or (8).

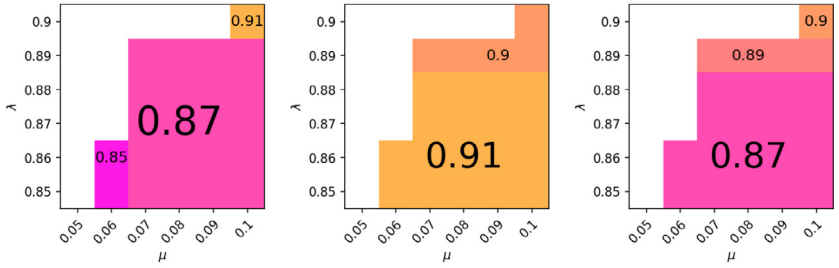
For the partitioning model, we use a similar procedure and the results can be found in Figs. 10 and 11. We can see from those figures that the conclusions are similar.



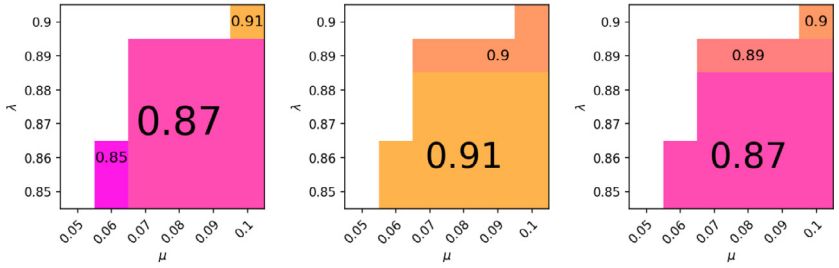
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{TPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{TPR}_c, c = 1, 2, 3$.

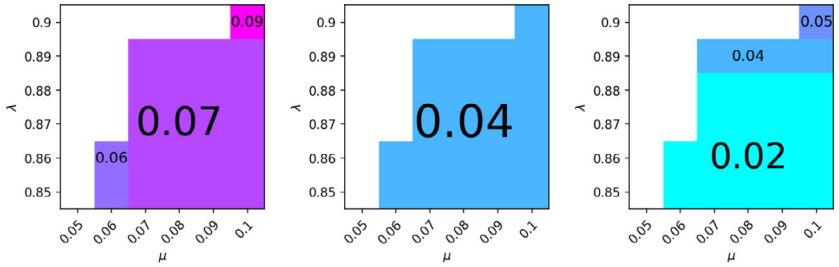


(c) $|\mathcal{N}| = 10^5$, $\mathrm{TPR}_c, c = 1, 2, 3$.

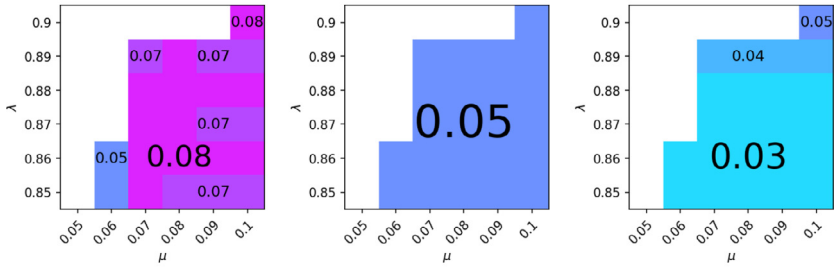


(d) $|\mathcal{N}| = 10^6$, $\mathrm{TPR}_c, c = 1, 2, 3$.

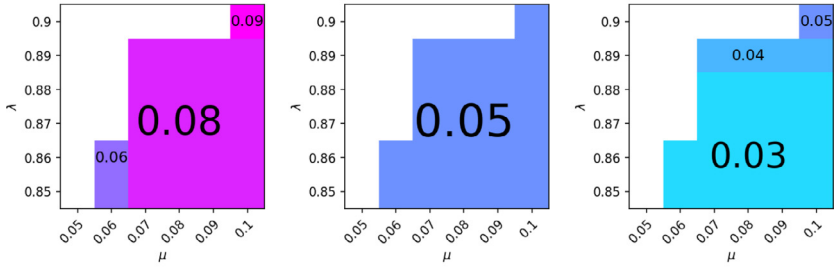
Fig. 8. For each cluster of the simulated data, the true positive ratio given by the covering model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



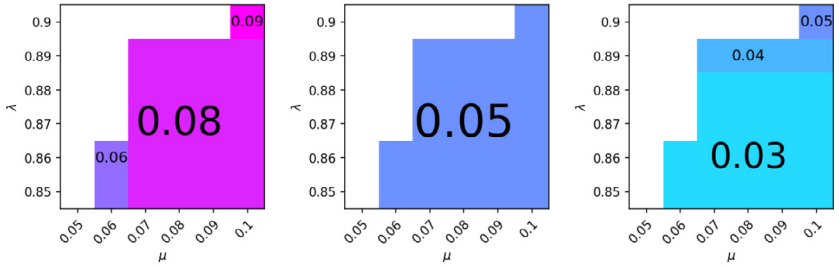
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{FPR}_c, c = 1, 2, 3$.

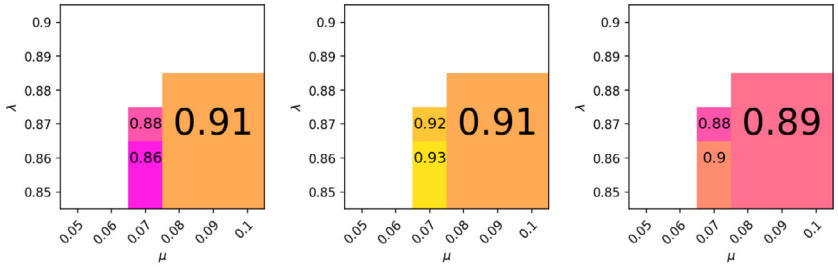


(c) $|\mathcal{N}| = 10^5$, $\mathrm{FPR}_c, c = 1, 2, 3$.

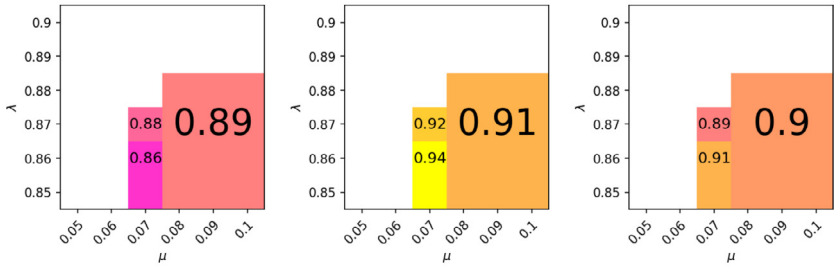


(d) $|\mathcal{N}| = 10^6$, $\mathrm{FPR}_c, c = 1, 2, 3$.

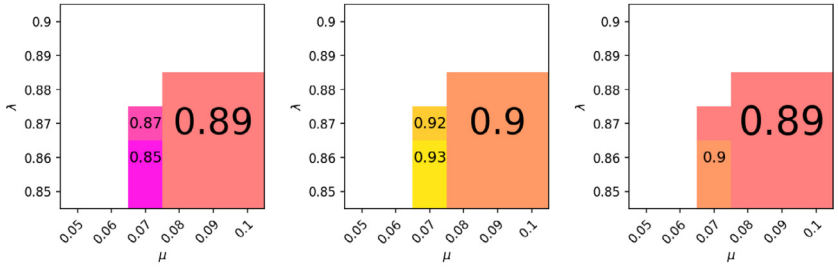
Fig. 9. For each cluster of the simulated data, the false positive ratio given by the covering model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



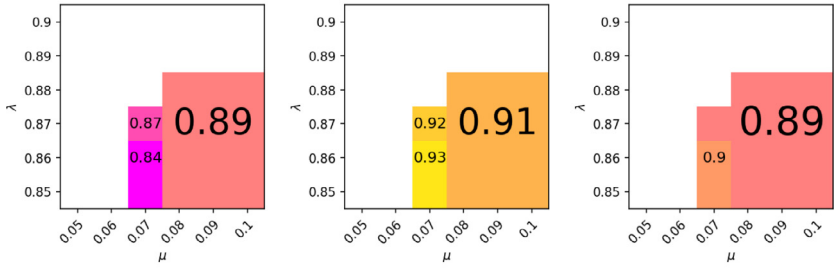
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{TPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{TPR}_c, c = 1, 2, 3$.

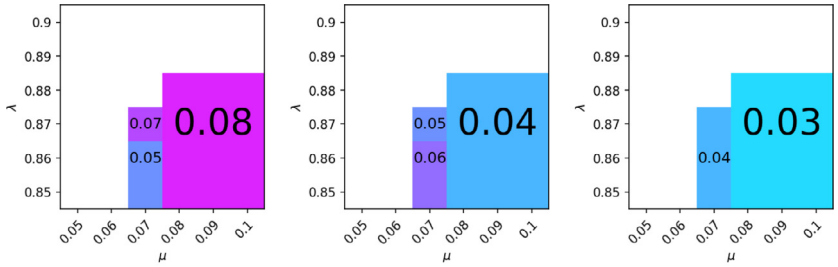


(c) $|\mathcal{N}| = 10^5$, $\mathrm{TPR}_c, c = 1, 2, 3$.

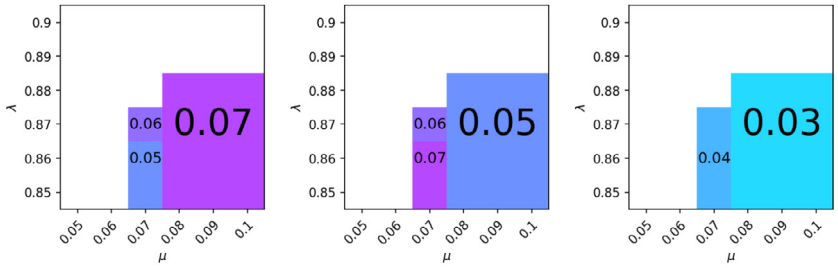


(d) $|\mathcal{N}| = 10^6$, $\mathrm{TPR}_c, c = 1, 2, 3$.

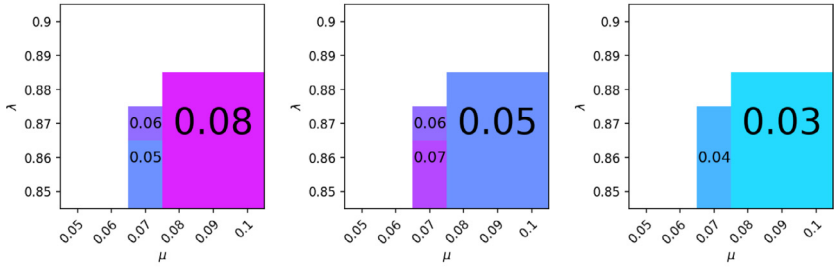
Fig. 10. For each cluster of the simulated data, the true positive ratio given by the partitioning model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



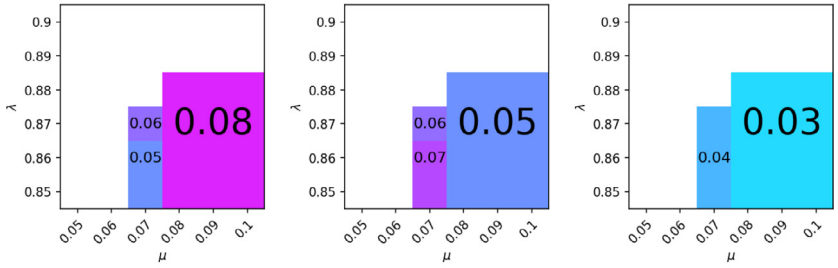
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(c) $|\mathcal{N}| = 10^5$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(d) $|\mathcal{N}| = 10^6$, $\mathrm{FPR}_c, c = 1, 2, 3$.

Fig. 11. For each cluster of the simulated data, the false positive ratio given by the partitioning model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.

5. Conclusions

In this paper, we have proposed a methodology to derive explanations for the clusters obtained from a Cluster Analysis procedure. The explanations are distance-based and defined as the set of individuals that are close to the so-called prototypes. To find explanations that are as accurate as possible, we select the prototypes that maximize the total number of true positive cases across all clusters and minimize the total number of false positive cases, while controlling the true positive rate as well as the false positive rate in each cluster. We have introduced two prototype optimization models, namely, the covering and the partitioning models. Both models can be formulated as MILPs. We illustrate the explanations provided by these models using both real-life data and simulated data.

There are two interesting lines of future research. The first one is to strengthen the mathematical optimization formulations provided in this paper, while the second one is to study the problem of building the clusters and explain them simultaneously.

CRediT authorship contribution statement

Emilio Carrizosa: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Kseniia Kurishchenko:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Alfredo Marín:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Dolores Romero Morales:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Acknowledgment

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214); FQM-329 and P18-FR-2369 (Junta de Andalucía, Spain); PID2019-110886RB-I00 (Ministerio de Ciencia e Innovación, Spain). This support is gratefully acknowledged. Part of this research was conducted while the third author, A. Marín, was on sabbatical at Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain.

References

[1] Baesens B, Setiono R, Mues C, Vanthienen J. Using neural network rule extraction and decision tables for credit-risk evaluation. Manage Sci 2003;49(3):312–29.
[2] Benítez-Peña S, Bogetoft P, Romero Morales D. Feature selection in data envelopment analysis: amathematical optimization approach. Omega (Westport) 2020;96:102068.
[3] Bertsimas D, King A. OR forum – an algorithmic approach to linear regression. Oper Res 2016;64(1):2–16.
[4] Carrizosa E, Molero-Río C, Romero Morales D. Mathematical optimization in classification and regression trees. TOP 2021;29(1):5–33.
[5] Carrizosa E, Romero Morales D. Supervised classification and mathematical optimization. Computers & Operations Research 2013;40(1):150–65.
[6] Freitas AA. Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter 2014;15(1):1–10.
[7] Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. Human decisions and machine predictions. Q J Econ 2018;133(1):237–93.
[8] Goodman B, Flaxman S. European union regulations on algorithmic decision–making and a "right to explanation". AI Magazine 2017;38(3):50–7.
[9] Lakkaraju H, Kamar E, Caruana R, Leskovec J. Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154.
[10] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016.
[11] Bénard C, Biau G, Da Veiga S, Scornet E. SIRUS: Making random forests interpretable. arXiv preprint arXiv:1908.06852.
[12] Carrizosa E, Nogales-Gómez A, Romero Morales D. Clustering categories in support vector machines. Omega (Westport) 2017;66:28–37.
[13] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Toward interpretable machine learning: transparent deep neural networks and beyond. arXiv preprint arXiv:2003.07631.
[14] Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications. ASA-SIAM Series on Statistics and Applied Probability. SIAM; 2007.
[15] Corral G, Armengol E, Fornells A, Golobardes E. Explanations of unsupervised learning clustering applied to data security analysis. Neurocomputing 2009;72(13):2754–62.
[16] Morichetta A, Casas P, Mellia M. EXPLAIN-IT: towards explainable ai for unsupervised network traffic analysis. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks - Big-DAMA '19; 2019.
[17] Gibert K, Conti D. On the understanding of profiles by means of post-processing techniques: an application to financial assets. Int J Comput Math 2016;93(5):807–20.
[18] Thomassey S, Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees. Decis Support Syst 2006;42(1):408–21.
[19] Ma R, Angryk RA, Riley P, Boubrahimi SF. Coronal mass ejection data clustering and visualization of decision trees. The Astrophysical Journal Supplement Series 2018;236(1):14.
[20] Bertsimas D, Orfanoudaki A, Wiberg H. Interpretable clustering: an optimization approach. Mach Learn 2021;110(1):89–138.
[21] Chen J, Chang Y, Hobbs B, Castaldi P, Cho M, Silverman E, et al. Interpretable clustering via discriminative rectangle mixture model. In: 2016 IEEE 16th International Conference on Data Mining (ICDM); 2016. p. 823–8.
[22] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[23] Davidson I, Gourru A, Ravi S. The cluster description problem - complexity results, formulations and approximations. Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc; 2018.
[24] De Koninck P, De Weerdt J, vanden Broucke SL. Explaining clusterings of process instances. Data Min Knowl Discov 2017;31(3):774–808.
[25] Balabaeva K, Kovalchuk S. Post-hoc interpretation of clinical pathways clustering using bayesian inference. Procedia Comput Sci 2020:264–73.
[26] Kauffmann J, Esders M, Montavon G, Samek W, Müller KR. From clustering to cluster explanations via neural networks. arXiv preprint:1906.07633.
[27] Aloise D, Hansen P, Liberti L. An improved column generation algorithm for minimum sum-of-squares clustering. Math Program 2012;131(1–2):195–220.
[28] Aloise D, Deshpande A, Hansen P, Popat P. NP-hardness of Euclidean sum-of-squares clustering. Mach Learn 2009;75(2):245–8.
[29] Grötschel M, Wakabayashi Y. A cutting plane algorithm for a clustering problem. Math Program 1989;45(1):59–96.
[30] Jain A. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010;31(8):651–66.
[31] Maldonado S, Carrizosa E, Weber R. Kernel penalized K-means: a feature selection method based on kernel K-means. Inf Sci (Ny) 2015;322:150–60.
[32] Rao M. Cluster analysis and mathematical programming. J Am Stat Assoc 1971;66(335):622–6.
[33] Seref O, Fan YJ, Chaovalitwongse WA. Mathematical programming formulations and algorithms for discrete-median clustering of time-series data. INFORMS J Comput 2014:160–72.
[34] Kaufmann L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley, New York; 1990.
[35] Carrizosa E, Martín-Barragán B, Romero Morales D, Plastria F. On the selection of the globally optimal prototype subset for nearest-neighbor classification. INFORMS J Comput 2007;19(3):470–9.
[36] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13:21–7.
[37] García S, Marín A. Covering location problems. In: Laporte G, Nickel S, Saldanha da Gama F, editors. Location Science. Springer International Publishing, Cham; 2019. p. 99–119.
[38] García S, Labbé M, Marín A. Solving large-median problems with a radius formulation. INFORMS J Comput 2011:546–56.
[39] Marín A, Pelegrín M. In: Laporte G, Nickel S, Saldanha da Gama F, editors. *p*-Median problems. Location Science Springer International Publishing, Cham; 2019. p. 25–50.
[40] Fortet R. Applications de l'algebre de boole en recherche opérationelle. Revue Française de Recherche Opérationelle 1960;4(14):17–26.
[41] Wagner J, Falkson L. The optimal nodal location of public facilities with price-sensitive demand. Geogr Anal 1975;7(1):69–83.
[42] Febrero-Bande M, Oviedo de la Fuente M. Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw 2012;51(4):1–28.
[43] Gurobi Optimization. Gurobi optimizer reference manual. 2020. URL http://www.gurobi.com.
[44] Python Core Team. Python: a dynamic, open source programming language. Python Software Foundation; 2015. URL https://www.python.org.