Measures of diversity and space-filling designs for categorical data

Cedric Malherbe[1], Emilio Dominguez, Merwan Barlier, Igor Colin[2], Haitham Bou-Ammar[3], Tom Diethe[1]
[1] Centre for AI, BioPharmaceuticals R&D, AstraZeneca, [2] LTCI, Télécom Paris, [3] Huawei Technologies

AstraZeneca

Summary

Problem How to measure the diversity of discrete sequences (e.g. biological and text data)? How to be create balanced training sets for such data?

Goal Design efficient algorithms to provide diverse sets of discrete sequences and provide algorithms to measure their diversity

Approach Relies on combinatorial optimization and greedy algorithms to create approximate algorithms

Problem statement

Diversity problem

Given a number $n \geq 1$ of points to select, we aim at constructing a subset $D_n = (x_1, \ldots, x_n)$ of n points in the boolean hypercube, denoted here by $= \{0,1\}^d$ for any dimension $d \geq 2$, called a design, that preserves the diversity of the space suitably. Since the notion of diversity does not admit a unique definition, we will focus on designs solutions to the following problems:

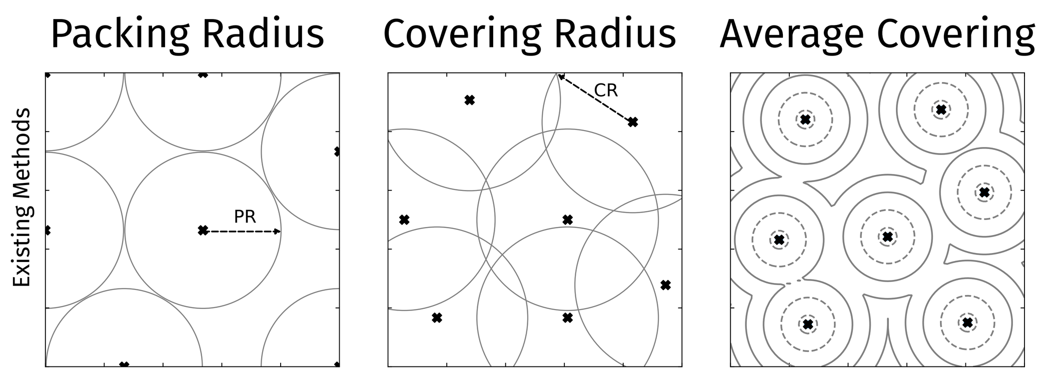
$$\text{find} \quad D_n^* := (x_1, \ldots x_n) \in^n \qquad \text{such that } \ell(D_n^*) = \min_{D_n \in^n} \ell(D_n) \tag{1}$$

where $\ell : ^n \to \mathbb{R}$ is a fixed measure of the diversity of a design D_n. To be more precise, we focus on creating designs optimizing three different diversity measures which are further defined below: the covering radius, the packing radius, and the average covering.

Contributions

- Three notions of diversity in categorical spaces: the average covering, packing, and covering radii,
- Theoretical results for the construction of optimal categorical designs
- Two novel approximation algorithms based on greedy schemes GRIPPR and GAC
- Experimental results validating the efficiency of the method

Measures of diversity



[0,0,0,0,0,0,0,0]	[0,1,1,1,0,0,1,0]	[0,1,1,1,0,0,1,0]
[0,0,1,1,1,1,1,1]	[1,1,0,0,1,0,0,1]	[1,1,0,0,1,0,0,1]
[1,0,0,1,0,1,0,1]	[0,0,1,0,1,0,0,0]	[0,0,1,0,1,0,0,0]
[1,1,1,1,1,0,0,1]	[1,1,0,0,0,0,1,1]	[1,1,0,0,0,0,1,1]
[1,1,1,1,0,1,1,0]	[0,1,0,1,0,1,0,1]	[0,1,0,1,0,1,0,1]

Figure 1: **Top:** 7 points that optimize the Packing Radius, Covering Radius, and Average Covering in the continuous space $\mathcal{X} = [0,1]^2$. The and radii are plotted in grey as well as the level sets of the function $x \mapsto d(x, D_n)$. **Bottom:** 5 points provided in this paper which optimize the same diversity metrics when $\mathcal{X}_8 = \{0,1\}^8$ is a categorical space.

Definition

(Packing Radius). Let $D_n = (x_1, \ldots, x_n) \in \mathcal{X}^n$ be any design of $n \geq 2$ points of the categorical space. Then, we define the packing radius of the design D_n as follows:

$$\text{PR}(D_n) := \min_{x_i \neq x_j \in D_n} \frac{d_H(x_i, x_j)}{2}.$$

Definition

(Covering Radius). Let $D_n = (x_1, \ldots, x_n)$ be any design of $n \geq 1$ points of the categorical space. Then, we define the covering radius of the design D_n over as follows:

$$\text{CR}(D_n) := \max_{x \in \mathcal{X}} d_H(x, D_n) = \max_{x \in \mathcal{X}} \min_{i=1 \ldots n} d_H(x, x_i).$$

Definition

(Average Covering). The average covering of a design $D_n = (x_1, \ldots, x_n)$ of $n \geq 1$ points of the boolean hypercube is defined as follows:

$$\text{AC}(D_n) := \mathbb{E}\left[d_H(X, D_n)\right] = \mathbb{E}\left[\min_{i=1 \ldots n} d_H(X, x_i)\right].$$

where $X \sim \mathcal{U}(\mathcal{X})$ is uniformly distributed over the space.

Algorithms for diversity

GRIPPR

Input: Dimensionality $d \geq 1$ of the categorical space, size $n \geq 2$ of the design

1. Set randomly the first design point $D_1 \leftarrow \{x_1\}$ where $x_1 \sim \mathcal{U}(\mathcal{X})$
3. For $t = 1, \ldots, n-1$:
Set

$$x_{t+1} \leftarrow \arg\max_{x \in} d(x, D_t)$$

Add the point to the design: $D_{t+1} \leftarrow D_t \cup \{x_{t+1}\}$
4. **Return** the design D_n

Theorem

The design D_n of GRIPPR satisfies:

$$\text{CR}_n^* \leq \text{CR}(D_n) \leq 2 \cdot \text{CR}_n^*$$

and

$$\frac{1}{2} \cdot \text{PR}_n^* \leq \text{PR}(D_n) \leq \text{PR}_n^*.$$

GAC

Input: Dimensionality $d \geq 1$ of the categorical space, size $n \geq 2$ of the design

1. Set randomly the first design point $D_1 \leftarrow \{x_1\}$ where $x_1 \sim \mathcal{U}(\mathcal{X})$
2. For $t = 1, \ldots, n-1$:
Get any point that greedily minimizes the average covering:

$$x_{t+1} \in \arg\min_{x \in \mathcal{X}} {}_{X \sim \mathcal{U}()}[d_H(X, D_t \cup \{x\})]$$

Set the novel design point: $x_{t+1} \leftarrow (\mathbf{x}_1^*, \ldots, \mathbf{x}_d^*)$
Add the point to the design: $D_{t+1} \leftarrow D_t \cup \{x_{t+1}\}$
4. **Return** the design D_n

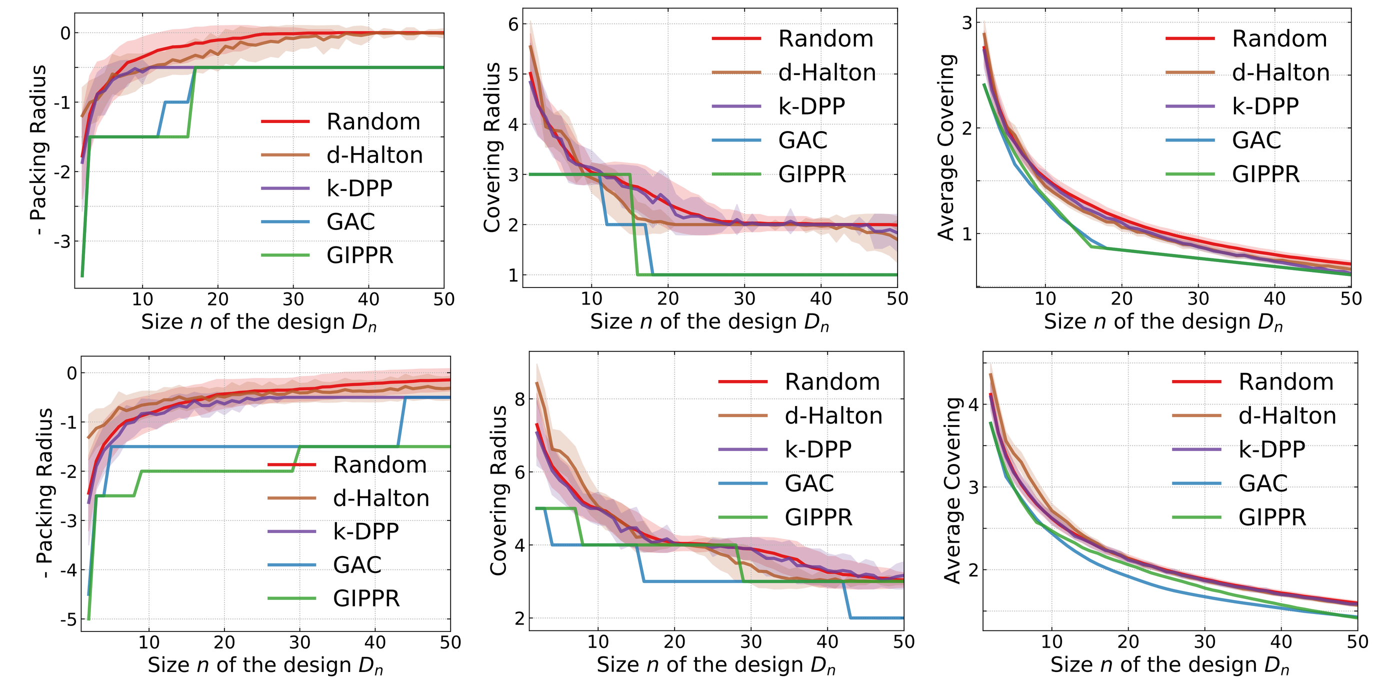
Theorem

The design D_n of GAC satisfies:

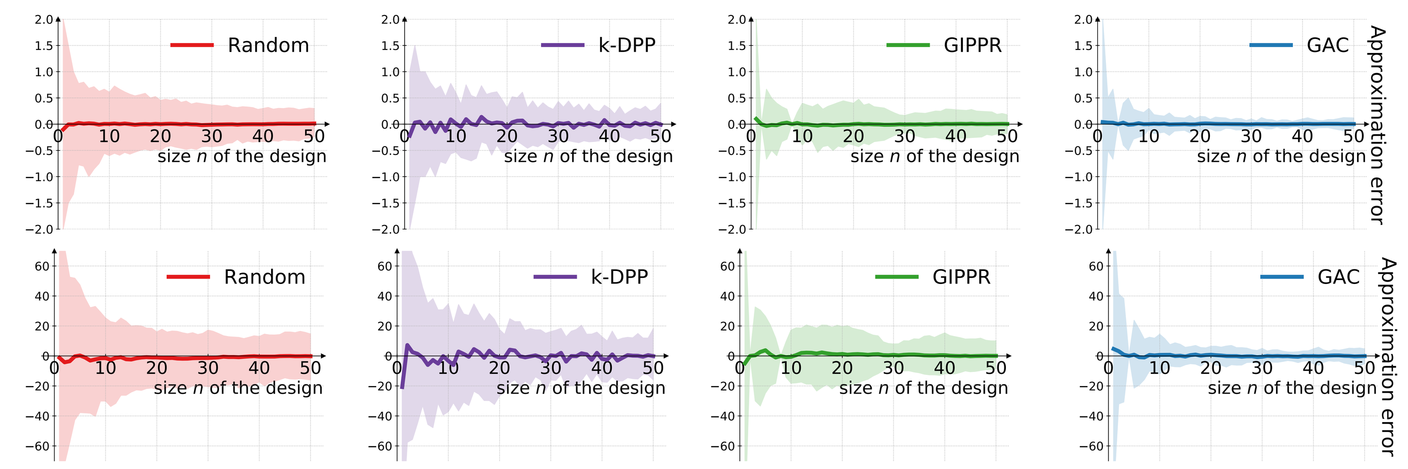
$$\frac{\text{AC}_1^* - \text{AC}_n^*}{2} \leq \text{AC}_1^* - \text{AC}(D_n) \leq \text{AC}_1^* - \text{AC}_n^*.$$

Empirical results

The graph displays the evolution of the diversity measures PR(D_n), CR(D_n) and AC(D_n) for different design sizes $n \in \{2, \ldots, 50\}$



The top line considers the case when $d = 7$ and the bottom line considers the case when $d = 8$. For each of the plots, lower is better.



The graphs display the average approximation error $\widehat{F}_n(D_n) - \mathbb{E}_{X \sim \mathcal{U}()}[f(X)]$ in bold for various design sizes $n \in \{1, \ldots, 50\}$, and the transparent colors represent the 90% and 10% quantile of the error computed over 100 runs with $d = 10$ for OneMax $f(x) = \sum_{i=1}^d \mathbb{I}\{x_i = 1\}$ (Top) and Harmonic $f(x) = \sum_{i=1}^d i^2 \mathbb{I}\{x_i = 1\}$ (Bottom).

References

Gomez, A. N., Pronzato, L., and Rendas, M.-J. Incremental space-filling design based on coverings and spacings: improving upon low discrepancy sequences. Journal of Statistical Theory and Practice, 15(4):1–30, 2021
Hammersley, J. M. Monte carlo methods for solving multi- variable problems. Annals of the New York Academy of Sciences, 86(3):844–874, 1960.
Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondr ak, J., and Krause, A. Lazier than lazy greedy. In Proceed- ings of the AAAI Conference on Artificial Intelligence, volume 29, 2015.
Pronzato, L. Minimax and maximin space-filling designs: some properties and meth- ods for construction. Journal de la Soci et e Franc aise de Statistique, 158(1):7–36, 2017
Gonzalez, T. F. Clustering to minimize the maximum in- tercluster distance. Theoret- ical computer science, 38: 293–306, 1985