OUTLIER DETECTION IN HIGH-DIMENSIONAL SPACES USING ONE-DIMENSIONAL NEIGHBORHOODS

Joris Falip, Frédéric Blanchard, Michel Herbin
joris.falip@univ-reims.fr

23 Janvier 2018

CReSTIC, Université Reims Champagne-Ardenne

CONTEXTE

"I´m going on an adventure!" - Bilbo Baggins

CONTEXTE DE LA PROBLÉMATIQUE

On retrouve le problème d'*outlier detection* dans de nombreux champs d'application :

- Détection d'intrusion, de fraudes
- Prétraitement de données
- Aide à la décision

L'ANALYSE MÉDICALE

Détecter les patients atypiques pour assister les décisions médicales.

Approche préliminaire basé sur des travaux antérieurs [1].

---

1. Herbin, M. et al., "Concept of Observer to Detect Special Cases in a Multidimensional Dataset.", I4CS, 2017.

L'ANALYSE MÉDICALE

Des difficultés inhérentes

- Pas de métrique pertinente a priori
- Pas de hiérarchisation des variables
- Interprétabilité
- Dimensionnalité [2]

2. Donoho, D. L. et al. "High-dimensional data analysis : The curses and blessings of dimensionality.", AMS Math Challenges, 2000.

OBJECTIFS

Solution attendue

- Visualisation et exploration
- Recommandation

Reposant sur la comparaison et structuration automatique [3]

3. Falip, J. et al. "Représentativité, généricité et singularité : augmentation de données pour l'exploration de dossiers médicaux.", EGC Atelier VIF, 2017.

OBJECTIFS

Solution attendue

- Visualisation et exploration
- Recommandation

Reposant sur la comparaison et structuration automatique [3]

Objectifs multiples

- Faciliter les stratégies médicales personnalisées
- Capitaliser les connaissances d'experts médicaux
- Assister les études épidémiologiques

---

3. Falip, J. et al. "Représentativité, généricité et singularité : augmentation de données pour l'exploration de dossiers médicaux.", EGC Atelier VIF, 2017.

MÉTHODOLOGIE

"What is this new devilry?" - Boromir

VUE D'ENSEMBLE

Établir un moyen de quantifier le caractère atypique :

- Adapté à la multi-dimensionnalité
- Permettant d'ordonner les patients
- Facilement explicable

ALGORITHME

Soit un ensemble de N éléments, définis sur D dimensions.

Etapes préliminaires

Pour chaque dimension d on calcule :

- Matrice de dissimilarité des éléments deux à deux
- Transformation des dissimilarités en rangs

On notera $Rank_e^d(n)$ le rang attribué par e à l'élément n sur la dimension d.

ALGORITHME

$Rank_e^d(n)$: rang attribué par ***e*** à l'élément ***n*** sur la dimension ***d***

$Knn^{d'}(n)$: ensemble des ***K*** plus proches voisins de ***n***

ALGORITHME

$Rank_e^d(n)$: rang attribué par ***e*** à l'élément ***n*** sur la dimension ***d***

$Knn^{d'}(n)$: ensemble des ***K*** plus proches voisins de ***n***

$$Rareness_e^d(n) = \frac{1}{K} \times \min(Rank_e^d(n), K) \tag{1}$$

ALGORITHME

$Rank_e^d(n)$: rang attribué par ***e*** à l'élément ***n*** sur la dimension ***d***

$Knn^{d'}(n)$: ensemble des ***K*** plus proches voisins de ***n***

$$Rareness_e^d(n) = \frac{1}{K} \times \min(Rank_e^d(n), K) \tag{1}$$

$$Rareness^d_{Knn^{d'}(n)}(n) = \frac{1}{K-1} \times \sum_{e \in Knn^{d'}(n)} Rareness_e^d(n) \tag{2}$$

ALGORITHME

$Rank_e^d(n)$: rang attribué par ***e*** à l'élément ***n*** sur la dimension ***d***

$Knn^{d'}(n)$: ensemble des ***K*** plus proches voisins de ***n***

$$Rareness_e^d(n) = \frac{1}{K} \times \min(Rank_e^d(n), K) \tag{1}$$

$$Rareness^d_{Knn^{d'}(n)}(n) = \frac{1}{K-1} \times \sum_{e \in Knn^{d'}(n)} Rareness_e^d(n) \tag{2}$$

$$Rareness(n) = \max_{\substack{d \in D \\ d' \in D}} \{Rareness^d_{Knn^{d'}(n)}(n)\} \tag{3}$$

ALGORITHME

Soit un ensemble de N éléments, définis sur D dimensions.

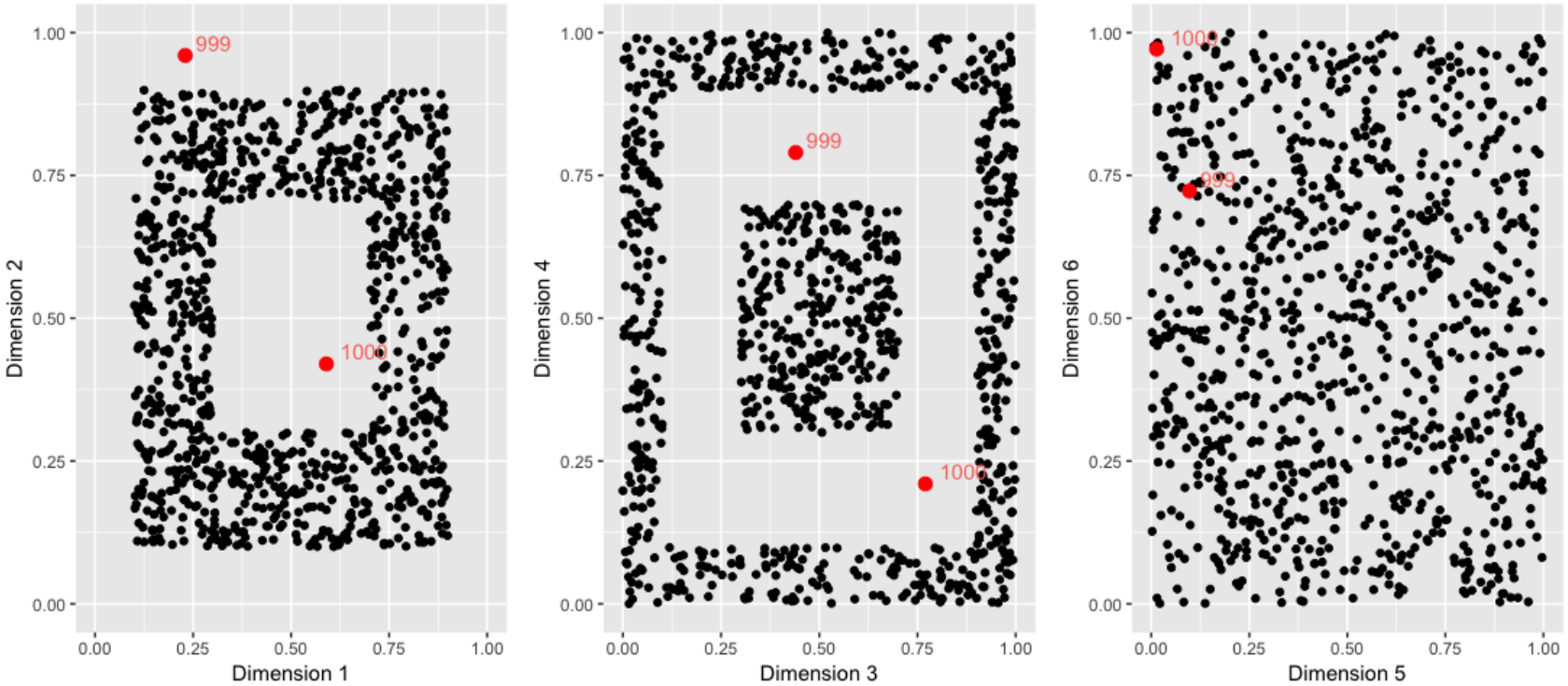
Complexité algorithmique :

- $N * D^2$ à l'heure actuelle
- $N * D$ en ne testant qu'un voisinage par dimension

ILLUSTRATIONS

"It's working!" - Peregrin Took

TOY EXAMPLE


999
1000
Dimension 2
Dimension 1
999
1000
Dimension 4
Dimension 3
1000
999
Dimension 6
Dimension 5
0.00
0.25
0.50
0.75
1.00

BENCHMARKS SYNTHÉTIQUES

Dix outliers dans chaque *dataset*.

Machine de test : 8 threads à 4GHz, 32Go RAM.

Table 1 – Temps d'exécution, en secondes

	Coeurs logiques			
Élements / **Dimensions**	1	2	4	8
1K / **10**	2	1	1	0.5
1K / **100**	40	26	13	12
10K/ **10**	157	80	45	38

CONCLUSION

"A wizard is never late." - Gandalf the Grey

EN QUELQUES MOTS

Résultats

- Première approche de l'*outlier detection*
- Interprétabilité des résultats
- Adéquation avec la haute dimensionnalité

EN QUELQUES MOTS

Résultats

- Première approche de l'*outlier detection*
- Interprétabilité des résultats
- Adéquation avec la haute dimensionnalité

Travaux futurs

- Comparaison à d'autres approches
- Étude de *datasets* réels, avec des experts
- Complexité espace/temps

OUTLIER DETECTION IN HIGH-DIMENSIONAL SPACES USING ONE-DIMENSIONAL NEIGHBORHOODS

Joris Falip, Frédéric Blanchard, Michel Herbin
joris.falip@univ-reims.fr

23 Janvier 2018

CReSTIC, Université Reims Champagne-Ardenne