CLUSTERING AVEC DES DONNÉES MANQUANTES ET DANS LE CONTEXTE DE CHIFFREMENT HOMOMORPHE

Code Master : MSQ689
Nom & Prénom : AZIZ Rezak
Encadré par : BOUZEFRANE Samia
AUDIGIER Vincent
BOUZAR Lydia
Promotion : 2021/2022

MOTS CLÉS

clustering, données manquantes, chiffrement homomorphe, vie privée

1. CONTEXTE ET MOTIVATION

- Large utilisation des méthodes de clustering.
- Problème de la vie privée.
- Problème des données manquantes.

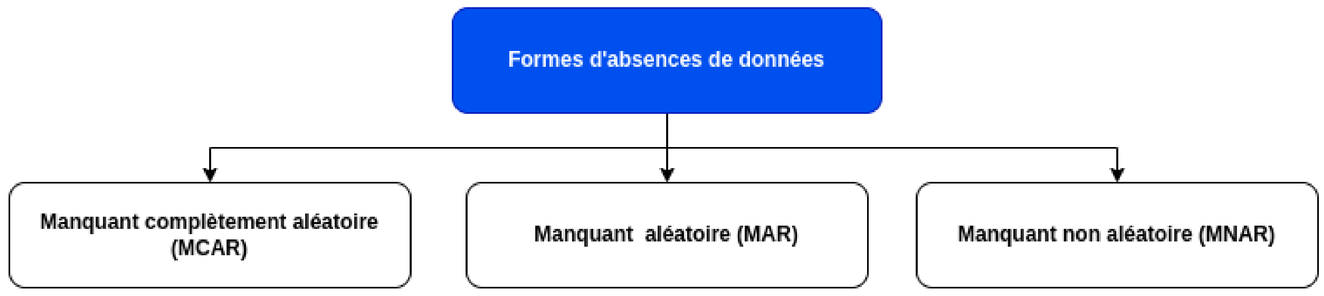
2. OBJECTIFS

- Etat de l'art sur le clustering avec des données manquantes.
- Etat de l'art sur le clustering dans le contexte de chiffrement homomorphe.
- Synthèse des méthodes existantes.

3. CLUSTERING

1. **Définition :**
   Regrouper des observations dans des clusters ou classes selon leurs similitudes.
2. **Types :**
   - Clustering Hard
   - Clustering Fuzzy ou flou
3. **Approches :**
   - Par partitionnement
   - Hiérarchique
   - Basée sur la densité
   - Basée sur une grille
   - Basée sur la distribution

4. DONNÉES MANQUANTES ET CLUSTERING



Gestion des données manquantes dans le clustering

- **Méthodes ad hoc :** Analyse de cas complet, imputation simple, ... => non adaptées
- **Méthodes directes :**

Travaux	Algorithmes	Solution proposée
[1]	k-means	Reformulation de problème d'optimisation en sautant les valeurs manquantes et le resoudre à l'aide d'un algorithme majoration-minimisation
[2]	k-means	Reformulation de problème d'optimisation et optimiser les 3 variables (matrice d'affectation, centres et données manquantes) en utilisant un algorithmes à 3 étapes
[3]	GMM	Solution appelée ignorable-GMM, modifier la formule de log-vraisemblance pour sauter les valeurs manquantes
[4]	GMM	Solution appelée k-CMM, prendre en compte les valeurs quantitative et qualitatives, imputation dynamique en utilisant les arbres de décision

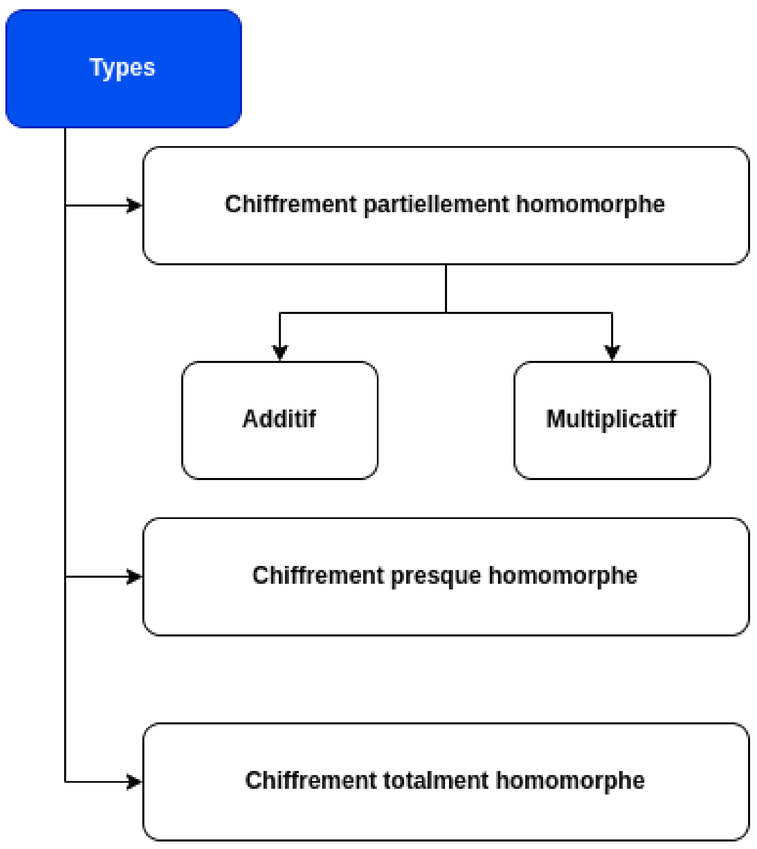
- **Imputation multiple :** Moins vu dans le cadre de clustering, cette méthode a été exploré dans le cadre de clustering par [5]

Validation des résultats :

- Validation interne, validation externe, sécurité.

5. CHIFFREMENT HOMOMORPHE

- Un chiffrement qui permet de faire des calculs sur des données chiffrées sans les déchiffrer.



Exemples :

- **PHE** Pallier(additif), RSA (multiplicatif)
- **SWHE :** BGN
- **FHE :** Gentry, BGV, CKKS, TFHE,...

6. CHIFFREMENT HOMOMORPHE ET CLUSTERING

Types de solutions: Solution collaborative (C), Solution individuelle (I)

Travaux	Algorithmes	Schémas	Type	Remarques
[6]	k-means	BGV	I	Comparaison et division au niveau de propriétaire des données, accepte des fuites d'informations
[7]	k-means	proposé	C	Calcul de cluster le plus proche au niveau de propriétaire des données
[8]	k-means	BGV	I	Utiliser un serveur de confiance qui déchiffre les distances et effectue les comparaison
[9]	k-means	Pallier	C	Utiliser une comparaison au niveau binaire, necessité de déchiffrement pour efectuer la multiplication
[10]	k-medoids	Pallier	C	Déchiffrement intermidiare pour la multiplication
[11]	k-means	Pallier	I	utilise la distance manhattan et guide l'algorithme avec une "Updatable distance matrix"

7. SYNTHÈSE

- Les algorithmes de clustering classique ne sont pas prévus pour gerer les données manquantes.
- Deux approches existent pour prendre en compte le problème de données manquantes dans le clustering: séparer l'imputation et le clustering, les approches directes.
- Le chiffrement homomorphe offre une solution pour la vie privée, mais ce dernier ne supporte que l'addition et la multiplication.
- Les points bloquants dans le clustering sont la comparaison, la division et le tri.
- Les solutions de clustering en utilisant le chiffrement homomorphes sont couteuses en termes de temps d'executions mais donnent des résultats presque équivalents aux solutions sans chiffrement.
- Aucun travail, à la limite de nos effors, combine les deux problématiques.

8. REFERENCES

[1] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 2016.

[2] Siwei Wang, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, and Jianping Yin. K-means clustering with incomplete data. 2019.

[3] Matthieu Marbac, Mohammed Sedki, and Tienne Patin. Variable selection for mixed data clustering: application in human population genomics. 2020.

[4] Duy-Tai Dinh, Van-Nam Huynh, and Songsak Sriboonchitta. Clustering mixed numerical and categorical data with missing values. 2021.

[5] Audigier and Niang. Clustering with missing data: which equivalent for rubin's rules? 2020.

[6] Anastasia Theodouli, Konstantinos A Draziotis, and Anastasios Gounaris. Implementing private k-means clustering using a lwe-based cryptosystem. 2017.

[7] Kai Xing, Chunqiang Hu, Jiguo Yu, Xiuzhen Cheng, and Fengjuan Zhang. Mutual privacy preserving k-means clustering in social participatory sensing. 2017.

[8] Gounaris Anastasios Sakellariou et al., Georgios. Homomorphically encrypted k-means on cloud-hosted servers with low client-side load. 2019.

REFERENCES (SUITE)

[1] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 2016.

[2] Siwei Wang, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, and Jianping Yin. K-means clustering with incomplete data. 2019.

[3] Matthieu Marbac, Mohammed Sedki, and Tienne Patin. Variable selection for mixed data clustering: application in human population genomics. 2020.

[4] Duy-Tai Dinh, Van-Nam Huynh, and Songsak Sriboonchitta. Clustering mixed numerical and categorical data with missing values. 2021.