

Introduction générale

L'intérêt de collecter et d'explorer de grandes quantités de données afin d'extraire une connaissance valable est devenu primordial pour les compagnies commerciales et les organisations gouvernementales, ce qui est la motivation du datamining. La tendance actuelle est que deux ou plusieurs organisations partagent leurs ensembles de données et les donnent comme entrée au processus du datamining afin d'avoir des résultats plus efficaces. Ceci a soulevé un réel problème de privacy du fait que la plupart de ces données concernent les individus et leurs informations personnelles. Le domaine de recherche très actif de la préservation de privacy en datamining a pour objectif d'extraire l'information utile à partir des données provenant de sources multiples, tout en préservant ces données contre la révélation ou la perte.

Notre travail [1] s'inscrit dans le domaine de la préservation de privacy en datamining en prenant le clustering comme une tâche du datamining. Le clustering sert à grouper ou classer des entités (points de données, items, échantillons,) décrites par des attributs selon certaines similitudes dans un contexte applicatif donné. Le processus consiste à former des groupes nommés clusters de telle sorte que les entités dans le même cluster soient similaires entre eux et dissimilaires aux entités des autres groupes. Plusieurs méthodes de clustering sont développées dans la littérature [2][3][4]. Ces méthodes sont capables de traiter différents types d'attributs et de découvrir des clusters de formes différentes. Parmi celles-ci, le clustering par partitionnement [2] qui regroupe les entités en des sous ensembles sans chevauchement, de telle façon que chaque entité (item) appartient exactement à un cluster, en se basant sur la minimisation d'une fonction objective. K-means [5] est l'un des algorithmes de cette méthode et le plus largement utilisé grâce à sa simplicité et son efficacité sur un large ensemble de données. La préservation de privacy dans l'algorithme k-means [1] est réalisée en protégeant les items (les données) d'un ensemble de données distribué sur plusieurs parties contre la révélation ou la perte lorsque l'algorithme k-means est exécuté. Ces parties participent dans l'exécution de l'algorithme k-means sans qu'aucune partie ne prenne connaissance des données (items) des autres parties. Ceci est réalisé dans le contexte du calcul multi - partie sécurisé [6][7][8].

Nous prenons connaissance alors d'un concept complexe et nouveau qui est Le calcul multi – partie sécurisé (Secure multi-party computation) traitant de la protection des données distribuées lors d'un calcul commun entre des parties différentes. Le calcul multi – partie sécurisé fait référence au problème général du calcul sécurisé d'une fonction à données distribuées. En 1982, A. Yao a initialement postulé le problème de la comparaison bipartite sécurisé et il a développé une solution prouvablement sécurisée [6]. Ceci a été étendu au calcul multi partie par O. Goldreich et al [7][8], qui ont développé le cadre (framework) formel du calcul multi partie sécurisé et ils démontrent que calculer confidentiellement (privately) une fonction est équivalent à la calculer d'une manière sécurisée.

Nous étudions et analysons les travaux de préservation de privacy dans l'algorithme k-means basés sur le calcul multi-partie sécurisé, nous classons ces algorithmes selon les modèles de distribution de données : vertical, horizontal et arbitraire car c'est le seul critère

qui affecte l'approche à entreprendre pour la préservation de privacy et présentons les limites et les points forts de chaque solution proposée, l'une par rapport à l'autre quand à la préservation de privacy. Ceci est réalisé en étudiant le niveau de protection des items distribués lors de l'exécution de l'algorithme k-means dans les différentes approches. L'objectif est de ressortir les besoins en privacy dans les différents algorithmes proposés et de tirer les meilleurs cas de préservation de privacy dans l'algorithme k-means.

Dans le premier chapitre, nous définissons le domaine de préservation de privacy en datamining, présentons les grands axes de recherches dans ce domaine, et aussi définissons les métriques d'évaluation des algorithmes de préservation de privacy en datamining et présentons aussi dans cette partie les modèles de distribution de données.

Dans le deuxième chapitre, nous nous consacrons au clustering comme technique du datamining, et à l'algorithme k-means comme algorithme très important de cette technique. Nous définissons et situons l'algorithme de k-means par rapport aux autres algorithmes de clustering.

Dans le troisième chapitre, nous décrivons le calcul multi-parti sécurisé, présentons ce domaine, définissons le modèle de sécurité utilisé par l'algorithme k-means qui est le modèle semi-honnête. Nous recueillons les primitives de sécurité les plus utilisées dans cet algorithme pour garantir la sécurité dans le modèle semi-honnête et nous les détaillons.

Dans le quatrième chapitre [1], nous étudions et analysons les différents algorithmes de préservation de privacy dans l'algorithme k-means selon une classification basée sur les modèles de distribution des données : vertical, horizontal et arbitraire. Dans chaque classe d'algorithmes (approche) nous analysons la protection des items dans les ensembles de données distribuées lors de l'exécution des différentes étapes de l'algorithme k-means. Ceci nous a permis de mesurer la préservation de privacy dans l'algorithme k-means dans le modèle de sécurité semi – honnête dans les différentes approches et de fixer les meilleurs cas de préservation de privacy dans l'algorithme k-means.