Appariement de données pseudonymisées

Catherine Quantin
DIM - CHU Dijon
catherine.quantin@chu-dijon.fr

Eric Benzenine
DIM - CHU Dijon
eric.benzenine@chu-dijon.fr

Maxence Guesdon
DIM - CHU Dijon
maxence.guesdon@inria.fr

Le Département d'Information Médicale (DIM) du CHU de Dijon doit régulièrement effectuer des appariements entre fichiers de santé.

La loi interdit la manipulation de données personnelles médicales non anonymes. Cependant, l'appariement de fichiers de données médicales personnelles est impossible si ces données sont anonymes au sens strict, puisqu'alors plus aucune information ne permet de relier un individu dans deux fichiers différents.

De plus, savoir si une donnée est indirectement identifiante par un tiers nécessite de connaître les informations dont peut disposer ce tiers. On a donc à faire à un "niveau d'anonymisation", qui n'est pas vraiment prévu par la loi.

Les sources de données à apparier sont donc pseudonymisées, en utilisant une fonction de hachage. Le hachage est une opération consistant à associer à une valeur d'un ensemble infini une valeur d'un ensemble fini. Deux valeurs en entrée peuvent avoir la même image (collision), cependant avec une probabilité très faible pour des valeurs d'entrée de tailles proches.

A partir de la valeur obtenue par hachage, il est impossible de "remonter" à la valeur d'origine, puisque par définition il en existe une infinité. Cependant, des attaques dites "par dictionnaire" sont possibles. Par exemple, il est possible d'appliquer la fonction de hachage à un ensemble de noms ou prénoms courants, pour obtenir une liste d'associations entre valeurs résultant du hachage et noms ou prénoms originaux.

Pour se protéger contre cette faille, qui entraîne un faible niveau d'anonymat des données, les données à hacher peuvent subir un traitement supplémentaire basé sur une transformation déterministe mais secrète. C'est ce qui est fait dans le logiciel Anonymat développé par le DIM, qui a reçu l'accord de la CNIL pour son utilisation sur des fichiers de données médicales.

Avec de telles sources de données pseudonymisées par une fonction déterministe, nous utilisons la méthode de Jaro pour effectuer des appariements probabilistes.

Cette méthode consiste à comparer chaque paire d'enregistrements issus des deux fichiers, pour lui affecter un score qui s'appuie sur le calcul du rapport de vraisemblance. Cette comparaison utilise certains champs des enregistrements à apparier. Ces champs sont choisis selon leur pouvoir discriminant. Une correspondance d'un même champ très discriminant entre deux enregistrements aura un poids fort dans le calcul final du score; au contraire, une correspondance selon un champ moins discriminant aura moins de poids.

La méthode de Jaro fournit le moyen de calculer les poids des champs et de fixer les seuils de probabilité d'appariement ou non. Chaque paire d'enregistrements se voit alors classée en "non appariée", "appariée" ou "indéfini". Les seuils pour classer automatiquement sont choisis selon le type d'étude (donc selon la sûreté d'appariement à atteindre), la qualité et la quantité des données. Une phase manuelle peut être nécessaire pour classer les paires qui n'ont pas été classées classées en "appariée" ou "non appariée" par la méthode.

Cette méthode d'appariement peut être utilisée dans différents buts:

- évaluer la qualité d'une base par rapport à un gold standard,
- combiner deux bases pour:
  - enrichir l'une ou l'autre de données supplémentaires,
  - effectuer des analyses non réalisables en ne disposant que de l'une ou l'autre.

Plusieurs applications pourront être présentées.