Sujet de stage Ingénieur, master 2

Construction d'un modèle bayésien naïf à 2 acteurs sous contrainte de respect de la vie privée

Stage de 5 mois, proposé par Orange Labs Lannion

Sujet

Le croisement d'informations de natures diverses détenues par différentes organisations a une forte valeur ; c'est un facteur de création de nouveaux usages. Le partage et la fouille de données personnelles nécessite cependant la mise en place de mécanismes de protection de la vie privée. Distribuer les données sur des sites géographiques différents est un moyen d'empêcher l'accès ou le contrôle de l'ensemble des données par un acteur unique.

On cherche à construire un modèle de Data Mining (en l'occurrence un classifieur) dans un contexte où les différentes sources de données nécessaires à sa construction sont distribuées sur différents sites. Les sources sont détenues par différentes parties qui ne souhaitent pas laisser l'accès aux enregistrements individuels.

Les stratégies que l'on rencontre dans la littérature pour résoudre ce problème sont en général très couteuses en opérations de calculs : elles consistent soit à faire appel à un tiers de confiance, soit à utiliser des protocoles cryptographiques. On souhaite développer une approche alternative pour calculer les paramètres du modèle.

Le modèle de fouille auquel on s'intéresse est le classifieur bayésien naïf dont les paramètres sont obtenus à partir des probabilités a priori des classes et des comptes conditionnels.

On peut envisager un premier scénario de distribution des données, dans lequel les variables explicatives et les classes sont détenues par des parties différentes qui ne souhaitent pas communiquer les identifiants qui permettraient de joindre les observations. Le calcul des comptes conditionnels n'est donc pas possible directement.

La solution que l'on souhaite développer consiste à encoder les données des différentes parties à l'aide d'outils comme les filtres de Bloom, puis à appliquer différentes opérations sur les filtres pour obtenir les comptes conditionnels.

Attendus du stage :

- Etat de l'art des solutions de la littérature de fouille de données respectueuses de la vie privée, avec des données distribuées
- Formalisation de la solution puis implémentation et évaluation sur différents jeux de données
- Rédaction d'un rapport et éventuellement d'une publication scientifique selon les résultats

Références :

- O. Papapetrou et al, *Cardinality estimation and dynamic length adaptation for Bloom filters.* Distributed and Parallel Databases December 2010, Volume 28, Issue 2-3, pp 119-156
- J. Vaidya. A survey of privacy preserving methods across vertically partitioned data, chapter 14 of book Privacy-Preserving Data Mining - Models and Algorithms. Advances in Database Systems, vol. 34. Springer (2008), pp 137-156, 2008

- Gandrade, R. Patel (2012). Privacy preserving naive Bayes classifier for horizontally distribution scenario using un-trusted third party. Journal of computer engineering, vol 7, issue 6, pp. 4-12

Profil recherché:

Bac+5 école d'ingénieur, master 2
Compétences en programmation nécessaires : maitrise d'un langage de script dédié à l'analyse de données (R, matlab, python)
Connaissances en statistiques, mathématiques et/ou apprentissage statistique
Connaissances en sécurité des systèmes d'information, protection des données personnelles sont un plus

Informations pratiques :

Lieu : Orange Labs Lannion, équipe PROF (Profilin et data Mining)
Le stagiaire sera intégré dans l'équipe de recherche sur le traitement statistique de l'information d'Orange Labs, directement en lien avec des problématiques opérationnelles du groupe Orange sur le CRM, l'Audience, le respect de la vie privée. Le stagiaire évoluera dans un contexte très recherche sur un sujet porteur.

Stage de 5 mois (mars 2015-juillet 2015), rémunéré (de l'ordre de 1000 euros /mois)

Contact : Françoise Fessant francoise.fessant@orange.com

Candidature : par email, avec des documents joints (CV, mémoire de M1 le cas échéant, etc.) exclusivement au format pdf.