Propositions de 3 sujets de thèse CIFRE en Informatique

Projet « Sosie Virtuel Projectif »

Introduction générale

La prise de décision est aujourd'hui une tâche ardue mais pour autant cruciale pour les différents acteurs économiques ou politiques. Pour faciliter cette dernière, certaines approches exploitent les données volumineuses, hétérogènes comme illustré dans (Athamena et al., 2018)[1]. Cette tâche, malgré cela, reste difficile selon les décisions qui doivent être prises. En effet, si l'on élargit le problème, avant de prendre une décision il est primordial de bien comprendre le contexte dans lequel la décision est prise et anticiper éventuellement le résultat de chacune d'entre elle. Or ce contexte n'est pas formellement explicité dans les données. En effet, par exemple, la/les relations entre les variables et les valeurs qu'elles prennent peuvent fortement dépendre de la situation dans laquelle elles existent. Le terme de « relation » est polymorphe et différents types de relations peuvent cohabiter sur le même jeu de données. Il peut par exemple exister des relations de corrélation mais également de dépendance. En effet, certaines variables ou groupes de variables peuvent avoir un impact sur un choix ou une décision et donc une valeur d'une autre ou d'un groupe de variable. Dans ce dernier cas, cet impact peut être vu également comme une explication de telle ou telle variable ou groupe de variable.

Dans le même temps, les approches actuelles arrivent à identifier les relations les plus présentes dans le jeu de données. Or cela ne suffit pas et il est important d'identifier les signaux faibles dans ces relations voire les relations implicites entre les données car ils peuvent être vus selon leur importance comme des « risques » liés à la future décision.

Par ailleurs, les relations ne sont pas statiques et bien souvent lorsque l'on étudie les données, ces relations évoluent au fil du temps. Il est donc important d'en rendre compte du décideur souhaitant anticiper la prise de décision.

L'objectif de ce projet en collaboration avec la société SolutionDataGroup est de prendre en compte ces problèmes dans ce que l'on qualifie de « sosie virtuel projectif ». Ce sosie a pour objectif de « comprendre » les données (données structurées ou non structurées) qui lui sont fournies et d'expliquer à l'utilisateur final, le décideur, les différents contextes dans lesquels interviennent les interactions entre les différentes variables. Cela permet en outre que le décideur puisse confronter sa propre vision avec celle proposée par le sosie.

[1] Athamena, Belkacem & Houhamdi, Z. (2018). Model for decision-making process with big data. Journal of Theoretical and Applied Information Technology. 96. 5951-5961.

La dimension « projective » se retrouve dans le fait que le projet doit non seulement traiter l'explicite dans les données mais doit s'approcher autant que possible de l'implicite (« du non dit ») dans ces données pour avoir une vision 360° des données.

Ce projet va considérer différentes applications de l'énergie à l'espace pour lesquelles différents jeux de données seront fournis. Dans certains cas ces données seront éclatées sur différents sites sans possibilité de les intégrer en totalité.

Dans ce projet l'accent sera mis sur les méthodes permettant d'extraire les interactions entre les données afin de dresser un rapport au décideur. Dans ce cadre, 3 sujets de thèses CIFRE sont proposés pour démarrage fin 2020. Ces sujets sont présentés dans les pages suivantes :

- **#1 - Découverte de relations explicites ou implicites dans les ensembles de données hétérogènes en grande dimension du Sosie Virtuel Projectif**
- **#2 - Le Federated-Learning au service du Sosie Virtuel Projectif**
- **#3 - Expliquer et valider par l'exemple un modèle de Sosie Virtuel Projectif**

Appel à candidatures :

Pour ces sujets de thèse, nous recherchons des candidat(e)s motivés, ayant une bonne ouverture d'esprit (le sujet touchant en effet à différents domaines et champs d'applications) ainsi que de bonnes compétences humaines et techniques dans les thématiques (IA, Machine Learning et Data Management) sous-jacentes à chaque sujet de thèse.

La localisation de la thèse sera sur Toulouse.

Les candidat(e)s intéressé(e)s par ces sujets de thèse doivent prendre contact au plus vite avec Max Chevalier (Max.Chevalier@irit.fr) et Arnaud Conrad-Bruat (aConrad-Bruat@solutiondatagroup.fr) en fournissant les documents suivants au format PDF :

- CV
- Lettre de motivation pour la/les thèses envisagées
- Copie du dernier diplôme (Master 2 ou Ingénieur ou équivalent)
- Notes de Licence ou Master (ou équivalent)
- Eventuelles lettres de recommandation pour une thèse

Sujet #1

Découverte de relations explicites ou implicites dans les ensembles de données hétérogènes en grande dimension du Sosie Virtuel Projectif

Une approche Multi Points de Vue

1. Contexte

Le sosie projectif est une intelligence artificielle construite pour notamment découvrir des phénomènes cachés au sein de masses de données. Nous entendons par phénomènes cachés, les relations explicites ou implicites, pouvant être de nature diverse (association, implication, etc), existantes dans un ensemble de données. Ces relations seront ensuite présentées à l'utilisateur final, c'est-à-dire le décideur, afin qu'il puisse les étudier, les explorer pour lui permettre de confirmer ou infirmer ses propres hypothèses quant à l'existence ou non de relations.

2. Problématique

Les masses de données posent deux problèmes importants à leur exploitation : hétérogénéité des données, et nombreuses caractéristiques des données (grandes dimensions). Il est impossible d'exploiter directement un ensemble de données hétérogènes en grande dimension pour l'élaboration du sosie projectif. Les approches existantes de l'apprentissage automatique exigent des pré-traitements afin d'obtenir des ensembles homogènes et réduits (basés sur une sélection de caractéristiques). Ces transformations peuvent induire des biais ou masquer certaines relations.

3. Objectif

L'objectif consiste à pouvoir directement explorer des masses de données hétérogènes pour découvrir des classes de relations explicites ou implicites possiblement cachées.

Nous proposons d'opter pour les méthodes de classification de données (*Clustering* [3]) qui constituent une famille de méthodes fondamentales dans l'apprentissage automatique (*Machine Learning*), qui est efficace pour identifier des catégories connues (approche supervisée) ou non (approche non-supervisée) dans les ensembles de données. Elle consiste à partitionner un ensemble de données en sous-ensembles disjoints en regroupant des données aussi similaires que possible, tout en assurant la plus forte dissimilarité possible entre les sous-ensembles.

Pour répondre au double enjeu de l'hétérogénéité et du grand nombre de dimensions, la stratégie choisie est le *Multi-View Clustering* [1, 2] Cette approche moderne constitue un moyen d'explorer des ensembles de données hétérogènes en fonction de différentes vues : une vue peut être basée sur une technique spécifique (approches par les distances, par densités ou noyaux de densités, etc), et exploiter un sous ensemble des caractéristiques des données.

L'utilisation de multiples vues permet de combiner différentes approches en fonction de différents sous-ensembles de caractéristiques, ceci afin de favoriser la découverte de relations cachées sur de grands ensembles hétérogènes.

Références

[1] Shaowei Wei, Jun Wang, Guoxian Yu, Carlotta Domeniconi, Xiangliang Zhang. Multi-View Multiple Clusterings Using Deep Matrix Factorization. AAAI'20, pp.6348-6355, 2000.

[2] Guoqing Chao, Shiliang Sun, Jinbo Bi. A Survey on Multi-View Clustering, arXiv:1712.06246, 2018

[3] Nabil El Malki, Franck Ravat, Olivier Teste. KD-means: Clustering Method for Massive Data based on KD-tree. DOLAP'20, pp.26-35, 2020.

Sujet #2

Le Federated-Learning au service du Sosie Virtuel Projectif

1. Contexte

Dans le contexte du sosie virtuel projectif (SVP), de multiples sources de données distribuées participeront à l'alimentation de cette intelligence artificielle. Ces données peuvent être : (i) distribuées verticalement où plusieurs partenaires détiennent les données d'un sous-ensemble de caractéristiques pour un même sujet ou (ii) distribuées horizontalement où les fournisseurs de données utilisent un même schéma de données mais ne couvrent qu'un sous ensemble des données.

2. Problématique

Pour résoudre les problèmes d'apprentissage sur les données dans un contexte distribué, le paradigme de Federated Learning (FL) [1], [3] est en pleine émergence. Ce paradigme repose sur l'implication des nœuds terminaux (NT) dans le processus d'apprentissage sans transfert de données. Le fonctionnement de prédilection consiste à un échange de paramètres d'apprentissage des NT vers un serveur agrégeant ces paramètres et les redistribuant vers les NT.

3. Objectifs

Dans cette thèse, nous nous appuyons sur le paradigme de FL pour développer deux méthodes afin de découvrir des relations implicites ou explicites qui pourraient exister dans les données distribuées sur lesquelles repose le SVP.

1. La première méthode se focalisera sur la détection de relations entre les données telles que par exemple les règles d'associations particulièrement dans un contexte de partitionnement vertical des données. Contrairement à d'autres approches (e.g. [4], [5]) de la littérature où le focus est l'aspect sécurité, notre méthode se doit générique puisque le SVP est souhaité multi-domaines et multi-applications.

2. La deuxième méthode se focalisera sur le contexte de partitionnement horizontal où les données concernent un seul domaine. Ces données souffrent des propriétés non-IID [3] induisant dans le même temps une répartition déséquilibrée entre les nœuds. La méthode poursuivie pourra notamment s'appuyer sur une orientation pluraliste [3] où nous ne chercherons pas à trouver un consensus entre les différents nœuds de données pour résoudre le problème de non-IID..

Une évaluation, de ces deux méthodes est attendue en mettant le focus sur la problématique des données se trouvant au croisement du partitionnement horizontal et vertical. L'évaluation répond à un double enjeu : efficacité des méthodes selon la notion de regret [7] appliquée au FL et acceptabilité générale des résultats selon les experts.

Références

[1] Qiang Yang, Yang Liu, Tianjian Chen, Yongxin Tong. Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology 10(2):1-19. 2019.

[2] Jiasi Weng, Jian Weng, Jilian Zhang, Ming Li, Yue Zhang, and Weiqi Luo. DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-based Incentive. IEEE Transactions on Dependable and Secure Computing,, 2018.

[3] Xu, Jie & Wang, Fei. Federated Learning for Healthcare Informatics. 2019

[4] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 639–644, July 2002.

[5] J. Vaidya and C. Clifton. Privacy preserving k-means clustering over vertically partitioned data. In The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2003.

[6] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. Journal of Machine Learning Research, 13(Jan):165–202, 2012.

Sujet #3

Expliquer et valider par l'exemple un modèle de Sosie Virtuel Projectif

1. Contexte

Pour permettre la confrontation de la vision du décideur avec le sosie projectif, c'est-à-dire lui permettre d'interpréter et de valider, le plus facilement possible, les conclusions restituées par le sosie, trois étapes principales doivent être poursuivies.

2. Objectifs

a) Identification des données représentatives et de leur valeur d'explicabilité :

La première étape devra permettre d'identifier, pour chaque conclusion du sosie, les données à remonter en priorité au décideur et qui devront être les plus représentatives parmi l'ensemble des données d'origine. Il faudra également s'assurer que ces données prioritaires ont une bonne valeur d'explicabilité permettant au décideur de valider la conclusion fournie par le sosie. En outre, les données étant amenées à évoluer au cours du temps, le temps devra être un élément à considérer pour, notamment, la remontée des données pertinentes.

b) Représentation visuelle avec le décideur :

Cependant, l'identification de ces données prioritaires ne suffit pas à assurer une bonne interprétation, par le décideur, des conclusions du sosie projectif. Ainsi, dans la deuxième étape, il s'agira de transformer, pour chaque conclusion du sosie, l'ensemble des données représentatives et leur valeur d'explicabilité en une représentation visuelle. Cette représentation visuelle doit permettre à l'utilisateur d'avoir une vue globale de ces données afin de faciliter leur interprétation.

c) Interaction avec le décideur et confrontation de nouvelles hypothèses :

Sur la base de la représentation visuelle et d'opérateurs de manipulation sur les données, il s'agira de permettre au décideur de :

- vérifier/valider les conclusions présentées par le sosie
- pouvoir tester de nouvelles hypothèses sur le modèle du sosie

A ce titre, le doctorant pourra s'inspirer des travaux de [1] et dont des optimisations ont été proposées dans [2] et [3].

Références

[1] Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res.11, 1–18 (Mar 2010)

[2] Ferrettini, G., Aligon, J., Soulé-Dupuy, C.: Explaining single predictions: A faster method. In: Chatzigeorgiou, A., Dondi, R., Herodotou, H., Kapoutsis, C.,Manolopoulos, Y., Papadopoulos, G.A., Sikora, F. (eds.) SOFSEM 2020: Theory and Practice of Computer Science. pp. 313–324. Springer International Publishing,Cham (2020)

[3] Ferrettini, G., Aligon, J., Soulé-Dupuy, C.: EImproving on coalitional prediction explanation. 24th European Conference on Advances in Databases and Information Systems, ADBIS 2020, August 25-27, 2020, Lyon, France. To Appear.