Proposition de sujet de thèse CIFRE 2016-2019 :

Détection d'anomalies dans les flux temps réels sol-bord de la SNCF

Mots clés :
· Apprentissage automatique · Fouille de données
· Détection d'anomalies · Modélisation de séries temporelles
· Supervision de flux temps réel · Trains communicants

Entités d'accueil :
– **Laboratoire LIRIS (UMR 5205)**, Domaine Scientifique La Doua, 43 bd 11 Novembre 1918, 69100 Villeurbanne
– **SNCF, DSI Voyageurs** - DD SI PF - Div. Architecture et Socles Communs, Tour Oxygène, 10-12 Bd Marius Vivier Merle 69393, Lyon Cedex 03.

1 Contexte

La Société Nationale des Chemins de fer Français (SNCF) produit et exploite dans son système d'informations une grande quantité de données hétérogènes récoltées en temps réel. Certaines d'entre elles, généralement liées à l'information voyageurs, sont en provenance du SI au sol et de ses applications opérationnelles, telles que les prochains départs ou passages de trains, les dessertes prévues, l'estimation des retards, les perturbations, la localisation au sol, etc. Mais elle dispose également d'informations en provenance du bord, à partir de trains dits « communicants », telles que les données de géo-localisation par GPS, les données de télé-maintenance, de suivi de mission, de comptage voyageurs, etc. La volumétrie de ces flux est variable et pourra aller, par exemple pour la géo-localisation, jusqu'à 200 messages par seconde.

Tous ces flux sont collectés en temps réel, agrégés, uniformisés et diffusés par des plate-formes dites « de médiation » en haute disponibilité. Ces dernières nécessitent une supervision de bout-en-bout, c'est-à-dire depuis les nombreux équipements émetteurs, variés et hétérogènes, jusqu'aux applications d'exploitation métier consommatrices de ces données, en passant par de multiples équipements intermédiaires. Ce type de supervision permet d'observer de nombreuses variations dans le trafic de données. Elles peuvent d'une part être causées par la dynamique des données récoltées (une perturbation du trafic, par exemple), et sont dans ce cas tout à fait normales. Cependant, elles peuvent aussi être non pas liées aux données observées, mais à l'infrastructure de collecte et de communication utilisée pour produire et faire transiter ces données. On parle alors de dysfonctionnements ou anomalies techniques de l'infrastructure, par opposition aux perturbations métier.

Ces anomalies concernent alors non pas les données métier circulant, mais les méta-données ou indicateurs relatifs aux flux observés (nombre de messages reçus par unité de temps, latence entre l'émission et la réception, etc.).

Lorsqu'une telle anomalie se produit, il est parfois difficile de s'en rendre compte, et le délai entre le début de l'anomalie et son constat peut être d'une journée entière, selon les cas. Quand à la détermination de la cause de l'anomalie, nécessaire à sa résolution, elle pourra s'étendre sur plusieurs journées.

Ce sont la détection automatique et l'analyse de ces anomalies liées à l'infrastructure informatique et de communication qui sont au cœur de cette thèse. Le travail consistera notamment au développement et à l'évaluation de techniques de détection d'anomalies appliquées aux flux temps réels sol-bord, et s'inscrit dans la continuité d'un stage de Master M2 dans lequel une première modélisation des flux et un algorithme de détection d'anomalies ont été mis en place. L'objectif de la thèse est d'obtenir un modèle dynamique complet capable de s'adapter aux changements de régimes dans les flux temps-réels d'une part, en limitant le nombre de faux positifs, et permettant d'autre part de prendre en compte un ensemble de connaissances métier comme le plan de transport théorique des trains et ses adaptations, le parc des trains communicants, ainsi que les relations de corrélations et de causalité éventuelles entre différents indicateurs.

La thèse s'effectuera dans le domaine et avec l'équipe « Trains Communicants » de la Direction Déléguée SI « Production Ferroviaire », au sein de la « DSI Voyageurs », dans l'EPIC « SNCF Mobilité » du groupe SNCF.

1.1 Rattachement au projet de supervision de bout-en-bout

Cette thèse est rattachée au projet de « supervision de bout-en-bout des flux temps réel sol-bord ». Il s'agit d'un projet initié par l'activité Transilien, ayant pour principal objectif d'assurer, en temps réel également, une supervision efficace, 24h/7J, des flux temps réel sol-bord, à tous les niveaux de la « chaîne » (émission, collecte, traitement et communication). Cette supervision implique d'être capable de détecter au plus tôt des situations anormales impactant ces flux et leur bonne exploitation, et ce, quelle qu'en soit la nature, le composant impacté, ou le composant responsable du problème.

L'objectif majeur de la thèse sera de contribuer à la détection et éventuellement l'analyse et la qualification des anomalies techniques de l'infrastructure informatique, mais avec pour seules données d'entrée l'observation des flux de messages (en termes de nombre de messages reçus, de latence des messages, et de certaines méta-données de contexte telles que les engins émetteurs, l'activité concernée, la position géographique, ou les numéros de train, de missions et de lignes concernés).

1.2 Objectifs applicatifs SNCF et positionnement vis à vis de la stratégie de l'entreprise

La principale problématique SNCF à laquelle ce travail contribuera est celle de l'*alerting* et de la réduction du délai de détection des problèmes. En effet, la détection d'anomalies devrait permettre de repérer au plus tôt une situation

anormale — voire même de la détecter avant que cette dernière n'impacte le métier — en générant une alerte adéquate dans les IHM des consoles de supervision. Cette alerte pourra soit déclencher une analyse plus approfondie par le *service desk* des exploitants, soit générer une nouvelle alerte escaladée vers les niveaux de supervision supérieurs ou vers les études.

Une autre problématique importante est celle de la pré-identification ou pré-qualification (automatique) de l'anomalie : la détection de l'anomalie, si elle est associée à une procédure de classification (par apprentissage ou par un mécanisme de type raisonnement à partir de cas) pourrait aider à pré-qualifier le type et le niveau de l'anomalie, pour mieux adresser les destinataires de l'alerte correspondante, et aider à sa résolution.

Cette détection d'anomalie et son éventuelle pré-qualification pourraient aussi contribuer, dans un contexte de retour d'expérience, à l'analyse *a posteriori* d'anomalies récurrentes présentant les mêmes caractéristiques (cycle et saisonnalité, durée, indicateurs et corrélations affectés, autres informations de contexte, etc.) sur un historique donné.

D'une manière générale, ce travail de thèse représentera une contribution importante dans les domaines de l'aide à la décision temps réel et de la supervision des flux temps réel, et présente donc un intérêt stratégique certain pour la SNCF. Ces sujets sont en effet particulièrement importants en cette période où l'information voyageurs temps réel et la production ferroviaire sont au centre des préoccupations. Ce travail s'articule d'ailleurs autour d'autres projets et axes de recherches identifiés au sein du réseau des experts internes de la SNCF (réseau Synapses), en particulier le *cluster* d'innovation et de recherche « Optimisation des Ressources et Exploitation » (ORE).

1.3 Disponibilité des données utiles à la thèse

Les plate-formes de médiation du domaine « Trains Communicants » traitent l'ensemble des flux temps réel sol-bord de l'activité Transilien, mais aussi des autres activités de transport de voyageurs comme Voyages (TGV et grandes lignes) ou TER.

Cela nous garantit un accès privilégié aux données des flux temps réel sol-bord. Les données de géo-localisation sont d'ores et déjà historisées, il conviendra pour les autres flux de s'assurer de disposer d'un historique suffisant pour permettre un apprentissage supervisé ou semi-supervisé.

2 État de l'art et verrous scientifiques

La détection d'anomalie est une application importante des deux domaines de l'apprentissage automatique et de la fouille de données en intelligence artificielle. Elle s'intéresse à la capacité d'un système à identifier des observations qui ne se conforment pas à une structure prévue, ou à un motif déjà présent dans un ensemble de motifs probables déjà observés. Alors que des données aberrantes (« outliers ») vont se traduire par une certaine rareté [1], des données anormales

vont pouvoir suivre des comportements plus complexes (poussées d'activité, délais variables de réapparition, variabilité de la probabilité d'apparition, etc.) [2]. Ainsi, des méthodes classiques de détection d'aberrations seront incapables de détecter de telles anormalités, ou bien au contraire détecteront de trop nombreux faux-positifs. Les méthodes de détection d'anomalies se divisent en trois principales familles :

- les méthode supervisées, qui vont se baser sur le pré-traitement d'un ensemble restreint de données sous la forme d'un étiquetage *a priori* (normal ou anormal), et qui utilisent principalement l'entraînement d'un classifieur destiné à classer les futures données dans l'une de ces deux classes [3,4] ;
- les méthodes non-supervisées, qui travaillent sur l'ensemble des données disponibles et partent du principe que les instances d'une même classe sont proches dans un espaces de représentation bien choisi et que donc cet espace est partitionnable — la classe « normale » étant a priori largement sur-représentée, il est facile de la distinguer de la classe « anormale » [5] ;
- les méthodes semi-supervisées, qui supposent que les instances d'entraînement ne sont étiquetées que pour une partie des données d'apprentissage — le modèle de représentation qu'elles construisent profite alors des avantages des deux familles ci-dessus (précision de l'étiquetage de l'approche supervisée et complétude de l'espace des données de l'apprentissage non supervisé) — c'est dans le cadre de ces deux dernières familles de méthodes que nous proposons de lever certains verrous scientifiques.

Dans le contexte de cette thèse, plusieurs questionnements seront explorés en utilisant les données mises à disposition :

1. Tout d'abord, la distinction normal/anormal sera élargie à un concept plus flou, qui permettra de détecter des comportements « suspicieux » ou « étranges ». Une mesure permettant d'évaluer la distance à la normalité devra être définie et normalisée, mais une méthode permettant d'adapter dynamiquement cette distance à l'évolution de l'infrastructure métier sera aussi proposée. À titre d'exemple, les experts analysant actuellement les données sont régulièrement confrontés à des événements prenant leur source dans le métier lui-même (une perturbation du trafic liée à une panne sur le réseau ferré, par exemple). Or ces événements sont actuellement difficilement dissociables d'autres pannes relatives à l'infrastructure de communication (un routeur bufferisant temporairement les données qu'il est censé faire transiter immédiatement, par exemple) sans prendre en compte le contexte de l'anomalie (présence ou absence d'une information connexe validant ou pas la panne du réseau ou la perturbation du trafic). Cette prise en compte du contexte de la dynamique d'apparition des événements sera un apport de cette thèse, ainsi que la formalisation de concepts d'expertise métier qui viendront enrichir le modèle de détection. Dans cette piste, il est aussi possible d'envisager l'utilisation de la fouille de motifs dynamiques dans des séries temporelles multiples (validés dans des concepts métiers) pour enrichir le modèle d'apprentissage, et s'adresser ainsi au cas de distributions complexes difficiles à

évaluer par échantillonnage.

2. Ensuite, une spécialisation de certaines méthodes actuelles de construction de modèles, caractérisées par leur application au cas particulier des séries temporelles hétérogènes multivariées (modèles à espace d'états), sera proposée. L'état de l'art actuel sur ces questionnements nous oriente vers la mise en œuvre de méthodes telles que l'inférence Bayesienne causale [6], et l'analyse de séries temporelles à l'aide d'outils de la dynamique des systèmes non linéaires [7]. La question de l'apprentissage ensembliste de tels modèles (combinaison optimale de classifieurs hétérogènes sous optimaux) devra notamment être abordée.
3. Enfin, une attention particulière sera portée sur la capacité de ces nouvelles méthodes à effectuer de l'apprentissage incrémental dans le temps, et à notamment remettre rapidement en question un ensemble d'instance observées dans le passé dans le cas ou une évolution de l'infrastructure rend de facto obsolète une partie du modèle appris (évolution du plan de transport ou du parc d'engins communicants, par exemple). Dans ce contexte, la détection et la prise en compte de changements de régime est une première piste possible, et la possibilité de permettre à un expert humain de guider l'apprentissage en fonction de ses connaissances métier en est une autre.

3 Jalons et organisation

Le travail de thèse sera articulé autour de différents jalons correspondant aux phases de la thèse et aux principaux axes développés, avec un jalon final sur la production du mémoire de thèse et la réalisation d'un prototype. À ces jalons s'ajouteront les productions d'articles scientifiques (et présentations associées en conférences), qui donneront également chacune lieu à une validation interne SNCF.

Pour l'instant, nous pouvons identifier les jalons potentiels suivants (toutefois ce lotissement évoluera en fonction des axes mis en avant de manière prioritaire) :

- amélioration de la détection d'anomalie, et application éventuelle à d'autres flux et à d'autres indicateurs ;
- prise en compte des changements de régime ;
- prise en compte de la corrélation et de la causalité entre indicateurs ;
- pré-qualification et/ou classification automatique ou semi-automatique des anomalies ;
- démonstration du prototype réalisé et présentation des résultats finaux.

À une échelle temporelle plus fine, le suivi de thèse sera assuré par les trois encadrants lors de réunions hebdomadaires. Un comité de pilotage regroupant des membres du laboratoire LIRIS et de la SNCF sera par ailleurs mis en place en début de première année. Son rôle sera d'effectuer des bilans trimestriels réguliers sur l'avancement des travaux. Enfin, la réinscription d'une année sur l'autre sera validée localement par la conseil de suivi des thèses du laboratoire LIRIS (qui fournira de plus au doctorant toute l'infrastructure de suivi classique du laboratoire), ainsi que par une commission annuelle de l'école doctorale InfoMath.

4 Informations pratiques et candidature

La date de démarrage prévue se situe entre Septembre et Novembre 2016. Les candidats intéressés doivent envoyer les documents suivants aux contacts listés dans la Section ci-dessous :

- une courte déclaration d'intérêt ;
- un CV détaillé ;
- une liste des cours et des évaluations scolaires des deux dernières années ;
- des lettres de recommandations potentielles.

La SNCF étant une entreprise dans laquelle le Français est l'unique langue utilisée, les candidats devront la maîtriser. Une connaissance de l'Anglais est bien-sûr aussi nécessaire.

Toutes les candidatures seront examinées au fur et à mesure de leur arrivée, et le poste restera ouvert jusqu'à ce qu'il soit pourvu.

Contacts

- **Denis Jouvin**, Architecte SOA du domaine Trains Communicants, Expert scientifique et technique du réseau SNCF SYNAPSES (denis.jouvin@sncf.fr), tél. 04 27 44 48 64, bureau 15-86
- **Serge Fenet**, Maître de conférences, Laboratoire LIRIS (serge.fenet@liris.cnrs.fr)
- **Christophe Rigotti**, Maître de conférences, HDR, Laboratoire LIRIS (christophe.rigotti@liris.cnrs.fr)

Bibliographie

[1] Charu C. Aggarwal, Philip S. Yu, Outlier Detection for High Dimensional Data. SIGMOD Conference 2001 : 37-46

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection : A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009).

[3] Joshi, M. V., Agarwal, R. C., and Kumar, V. 2001. Mining needle in a haystack : classifying rare classes via two-phase rule induction. In Proceedings of the 2001 ACM SIGMOD international conference on Management of data. ACM Press, New York, NY, USA, 91–102.

[4] Vilalta, R. and Ma, S. 2002. Predicting rare events in temporal domains. In Proceedings of the 2002 IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 474.

[5] Keogh, E., Lin, J., and Fu, A. 2005. Hot sax : Efficiently finding the most unusual time series subsequence. In Proceedings of the Fifth IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 226–233.

[6] Pearl, Judea. 2000. Causality : Models, Reasoning, and Inference. Cambridge University Press.

[7] Huanfei, Kazuyuki, Luonan. 2014. Detecting Causality from Nonlinear Dynamics with Short-term Time Series. In Nature.

[8] Hsu, Srivastava. 2009. Diversity in combinations of heterogeneous classifiers. Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD, Lecture Notes in Computer Science, vol. 5476, 2009, Springer, Berlin, Heidelberg, 923–932.