Projet HERELLES : Hétérogénéité des données - Hétérogénéité des méthodes : Un cadre collaboratif unifié pour l'analyse interactive de données temporelles

Thi-Bich-Hanh Dao*†[1]

[1]Laboratoire d'Informatique Fondamentale d'Orléans – Université d'Orléans : EA4022, Institut National des Sciences Appliquées - Centre Val de Loire : EA4022 – France

Résumé

La nature variée des capteurs et sources de production des données temporelles entraîne une hétérogénéité forte tant en formats, volumes, qualités qu'en richesses d'information. Face à cette surabondance de données temporelles, arrivant de façon quasi-continue, la phase d'étiquetage de l'apprentissage supervisé ne peut plus être assurée par les experts, car trop fastidieuse et chronophage. Pour pallier ce manque, l'expert doit pouvoir exploiter d'autres sources d'information. Il peut s'agir de données partiellement étiquetées, de connaissances formalisées, de contraintes sur les données ou sur les résultats. Parallèlement, les méthodes aptes à analyser ces données sont elles aussi nombreuses. Combiner ces données et méthodes semble de fait indispensable. Ainsi, des approches telles que le boosting, l'ensemble clustering ou le clustering collaboratif tirent parti de la complémentarité entre différentes méthodes, chacune avec ses propres biais et sa propre stratégie d'analyse mais capable de traiter ses propres données de façon privilégiée. Malheureusement, ces approches se limitent majoritairement à la combinaison d'algorithmes partageant le même paradigme (e.g. supervisé ou non supervisé) ce qui réduit leur intérêt. Notre première hypothèse (H1) est qu'une approche basée sur la collaboration de méthodes multiparadigmes permettra de bénéficier de façon accrue de la complémentarité des méthodes et des données.
Néanmoins, face à la complexité des données temporelles et des phénomènes étudiés, définir toutes les informations nécessaires à une analyse de qualité est fastidieux et chronophage voire impossible. En effet, dans de nombreux domaines, de par la nouveauté de telles données, les champs sémantiques appropriés pour qualifier ou contraindre les données peuvent n'avoir pas encore été parfaitement définis. Ils doivent alors être découverts progressivement pendant l'exploitation et l'analyse des données à travers un cycle d'interactions entre l'expert et le système d'apprentissage. Cette interaction vise à réduire le fossé entre les résultats produits par les algorithmes et les intuitions thématiques de l'expert et rendre les résultats plus compréhensibles pour celui-ci, donc plus aisés à associer à une sémantique du domaine d'application. Notre deuxième hypothèse (H2) est qu'une méthode active (incrémentale et interactive) dans laquelle l'utilisateur injecte à la volée des informations en fonction de l'avancement de l'analyse limitera fortement son implication directe et permettra une sémantisation plus aisée.

*Intervenant

†Auteur correspondant: thi-bich-hanh.dao@univ-orleans.fr

Cette interaction met en jeu des savoir-faire et des connaissances propres au domaine d'application et à l'expert. Ces savoir-faire et connaissances ainsi explicités sont des biens précieux qui doivent pouvoir être réutilisés. Notre troisième hypothèse (H3) est que les connaissances explicitées et les savoir-faire mis en œuvre lors de l'interaction doivent être capitalisés et analysés pour améliorer les usages futurs du système.

Enfin, les données temporelles présentent une forte hétérogénéité tant dans la fréquence que la nature de l'acquisition. Par exemple, en observation de la Terre, une série d'images satellitaires à très haute fréquence temporelle offre une vision quasi-continue, qualifiable de chronométrique, du phénomène analysé. Au contraire, avec les textes et documents Web qui rapportent majoritairement des faits ponctuels (décrets, inaugurations...) le phénomène est vu de façon discrète, offrant une vision plus chronologique. Cette complémentarité permet d'étudier des dynamiques temporelles variées et souvent imbriquées. Notre quatrième hypothèse (H4) est que combiner des données temporelles offrant des visions complémentaires permet de profiter au maximum de leur hétérogénéité plutôt que de la subir.

En rupture avec les approches actuelles basées chacune sur un seul paradigme d'analyse, le projet HERELLES propose de :

- Définir une architecture générique permettant de faire collaborer des méthodes multiparadigmes travaillant potentiellement sur des données différentes et de définir les conditions optimales de son utilisation (H1) ;
- Développer des mécanismes d'interaction avec l'utilisateur lui offrant la possibilité d'injecter de nouvelles informations et de réduire le fossé sémantique entre les résultats et les intuitions de l'expert (H2);
- Proposer des méthodes d'extraction et de capitalisation des connaissances directement ou indirectement produites lors de l'utilisation de la méthode collaborative active (H3) ;
- Mettre en œuvre et valider la proposition dans le cadre de l'analyse de séries temporelles hétérogènes (H4).

Consortium : ICUBE, AgroParisTech, GREYC, LIFO, TETIS