Structuration semi-supervisée des données complexes*

Marian-Andrei Rizoiu
Laboratoire ERIC
Université Lumière Lyon 2
Lyon, France
Marian-Andrei.Rizoiu@univ-lyon2.fr

Résumé

L'objectif de la thèse est d'explorer la façon dont les données complexes peuvent être analysées en utilisant des techniques d'apprentissage non-supervisé, dans lequel de l'information supplémentaire est injectée pour guider le processus exploratoire. A partir de problèmes spécifiques, nos contributions prennent en compte les différentes dimensions des données complexes : leur nature (*e.g.*, image, texte), l'information additionnelle attachée aux données (*e.g.*, étiquettes, structure, ontologie de concepts) et la dimension temporelle. Le travail de recherche réalisé dans le cadre de cette thèse porte sur deux grandes problématiques : introduire des connaissances sémantiques dans la représentation des données et prendre en compte le temps dans le processus d'apprentissage.

1 Cadre général

Le contexte général de la recherche effectuée dans cette thèse se situe à au croisement des domaines de l'**analyse de données complexes** et du **clustering semi-supervisé**. Le projet de recherche sous-jacent à cette thèse a été construit progressivement, au travers d'une relation dialectique entre la théorie et la pratique. Les projets de recherche dans lesquels j'ai été impliqué au fil de cette thèse ont soulevé plusieurs problèmes spécifiques, qui ont nécessité souvent de traiter des données complexes (données hétérogènes de différentes natures, *e.g.*, texte, image) et l'intégration d'information supplémentaire dans le processus d'apprentissage.

Les différentes contributions de cette thèse sont illustrées et structurées de manière conceptuelle par la figure 1. Cette structuration repose d'une part sur la translation des données de natures différentes dans un espace de représentation capable de capturer la sémantique des données, et d'autre part l'injection de connaissances externes dans les algorithmes d'apprentissage non-supervisés.

1.1 L'analyse des données complexes

Il est difficile de traiter efficacement les données complexes [8], celles-ci étant de nature variée (*i.e.*, texte, image ou audio/vidéo) et étant profondément hétérogènes

*Thèse préparée sous la direction de Stéphane Lallich et Julien Velcin, professeurs à l'Université Lumière Lyon 2.

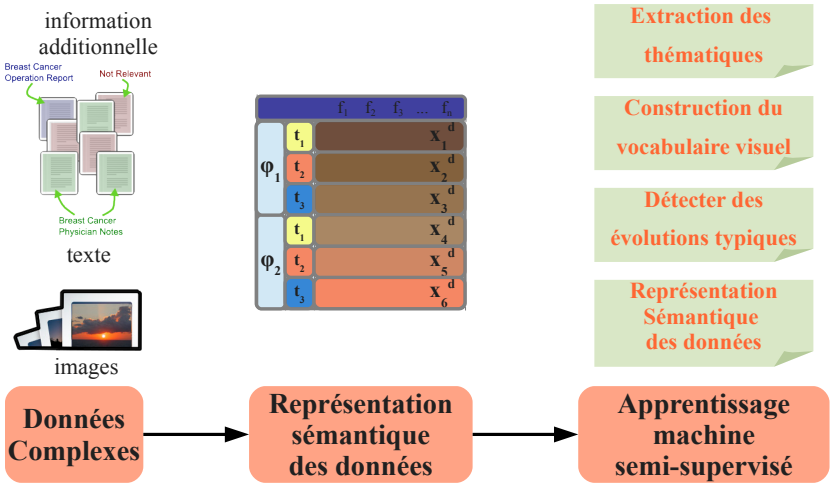

FIGURE 1 – Organisation conceptuelle du travail réalisé dans cette thèse.

puisque pouvant provenir de sources multiples. Les données complexes sont souvent temporelles, puisque l'évolution des entités dans le temps peut être enregistrée. De plus, des informations additionnelles sont souvent attachées aux données, sous la forme d'annotations d'experts, de structure des documents inter-connectés ou de bases de connaissances librement accessibles (*e.g.*, des ontologies comme DBpedia [1]).

Une méthode simple pour traiter les données complexes de différentes natures est de les représenter dans un espace numérique et d'appliquer des algorithmes classiques d'apprentissage automatique. La plupart des algorithmes d'analyse de données ont été développés pour utiliser des données décrites dans cet espace de représentation. Chaque document est représenté par un vecteur multidimensionnel, où chaque dimension correspond à une variable prédéfinie. Le défi actuel consiste à représenter les données de nature différente dans un espace numérique capable de capturer l'information sémantique présente dans le format natif tout en utilisant efficacement de l'information externe pour enrichir la sémantique de cet espace.

Parmi les méthodes permettant de prendre en compte l'information externe, nous avons privilégié les approches issues du clustering semi-supervisé que nous détaillons dans la section suivante.

1.2 Clustering semi-supervisé

Introduire des connaissances expertes partielles dans un algorithme du clustering relève du domaine du clustering semi-supervisé. Contrairement à l'apprentissage semi-supervisé, où l'accent est mis sur le traitement des données manquantes ou insuffisantes dans les algorithmes supervisés, le clustering semi-supervisé est utilisé lorsque que la quantité de supervision est tellement faible ou partielle qu'il est impossible d'appliquer des techniques supervisées. Les connaissances *a priori* se présentent soit sous la forme d'étiquettes de classe, soit sous la forme de contraintes sur des paires d'individus [7]. Elles sont ensuite utilisées pour guider le processus de clustering dans l'espace des solutions. Nous utilisons des techniques issues du clustering semi-supervisé pour modéliser et utiliser à la fois les connaissances additionnelles attachées aux données complexes et la dimension temporelle des données.

2 Contributions de la thèse

Les principales contribution de cette thèse sont articulées entre elles autour des deux principales problématiques à la base de notre recherche : construire un espace de représentation capable de capturer la sémantique sous-jacente aux données et prendre en compte l'aspect temporel des données. Ces deux problématiques de recherche se traduisent par des lignes directrices que l'on retrouve au travers notre travail : (a) obtenir des résultats facilement humainement interprétables, (b) plonger les données de différentes natures dans des espaces numériques capables de capturer la sémantique et (c) construire des algorithmes et méthodes prenant en compte des informations sémantiques et la dimension temporelle.

Par la suite, nous détaillons les contributions les plus importantes de notre travail de recherche.

Détecter les évolutions typiques Une de nos principales problématiques de recherche est de tirer parti de l'information temporelle dans le processus de regroupement non-supervisé. Dans [5], nous détectons les évolutions typiques des entités en proposant une nouvelle mesure de dissimilitude temporelle et une fonction de pénalité inspirée de la loi normale. La fonction de pénalité est utilisée avec des techniques de clustering semi-supervisé et a pour but d'encourager la segmentation contiguë des observations correspondant à une entité. Nous proposons un nouvel algorithme de clustering temporel, appelé TDCK-Means, qui crée une partition de clusters cohérente à la fois dans l'espace multidimensionnel et dans l'espace temporel.

Utiliser la sémantique des données pour améliorer l'espace de représentation Comme présenté dans la section 1.1, le traitement des données complexes de natures différentes (*e.g.*, image, texte) se résume habituellement à représenter les données dans un espace numérique et à appliquer des algorithmes classiques d'apprentissage. Nous considérons crucial d'améliorer cet espace de représentation des données pour prendre en compte les relations sémantiques issues du jeu de données. Nous construisons, en utilisant des algorithmes non-supervisés, de nouveaux attributs qui sont plus adaptés pour décrire l'ensemble des données tout en étant compréhensibles pour un utilisateur humain. Nous proposons dans [6] deux algorithmes pour construire de nouveaux attributs sous la forme de conjonctions d'attributs initiaux ou de leurs négations. Les attributs ainsi générés sont moins corrélés entre eux et mettent en valeur les relations sémantiques cachées entre les individus.

Améliorer la représentation des images en utilisant une construction semi-supervisée du vocabulaire visuel Une des façons les plus souvent utilisées pour traduire les images de leur format natif vers un format numérique est la représentation "sac-de-mots-visuels" (en anglais "bag-of-features"). Dans notre travail concernant les images, nous utilisons des connaissances expertes, sous la forme d'annotations, dans le processus de construction de la représentation numérique, à l'aide de techniques issues du clustering semi-supervisé. Nous proposons deux approches : dans la première nous construisons un vocabulaire visuel adapté pour décrire chaque objet qui apparaît dans la collection d'images, tandis que le second porte sur le filtrage des points d'intérêt qui ne concernent pas l'objet en cause.

Analyser des données textuelles : extraire et évaluer les thématiques Les données textuelles peuvent être transformées en format numérique en utilisant une représentation "sac-de-mots". Une fois cette représentation mise en place, les thématiques extraites des textes peuvent être utilisées, par exemple, pour la construction automatique d'ontologies de concepts [4]. L'extraction des thématiques peut aussi être amélioré en utilisant des informations supplémentaires. Dans [3], nous montrons comment une hiérarchie de concepts peut être utilisée pour évaluer les thématiques extraites en utilisant des approches statistiques (*e.g.*, LDA [2]).

3 Conclusion, travaux en cours et perspectives.

Les travaux effectués dans le contexte de cette thèse se situent au croisement de l'**analyse de données complexes** et du **clustering semi-supervisé**. Nous étudions comment les données de différentes natures peuvent être traitées tout en tenant compte de la dimension temporelle et de l'information supplémentaire attachée aux données. Les problématiques traitées dans cette thèse sont très vastes et soulèvent de nombreuses perspectives de travail. Les futurs travaux incluent une meilleure intégration des différentes approches proposées et de nos deux problématiques de recherche.

3.1 Travaux en cours

Actuellement nous travaillons sur l'extension et l'amélioration de nos approches. Nous développons une extension de notre algorithme de clustering temporel pour introduire une construction simultanée d'une structure de graphe entre les clusters obtenus. Nous sommes aussi intéressés d'améliorer notre algorithme de construction de attributs pour prendre en compte la dimension temporelle, en plus de la sémantique du jeu de données.

Une autre perspective est de généraliser l'utilisation de nos approches à d'autres types de données. Par exemple, nous sommes en train d'adapter notre algorithme de clustering temporel à la détection de rôles dans les réseaux sociaux en ligne.

Travaux appliqués

Les aspects théoriques de la thèse ont été développés conjointement avec la réalisation de prototypes. L'aboutissement de ces prototypes est `CommentWatcher`, une plateforme libre dédiée à l'analyse des discussions en ligne, plus précisément, des forums. Construite comme une plate-forme web, `CommentWatcher` dispose d'un module de récupération automatique des forums en utilisant une architecture modulaire, d'un module d'extraction des thématiques à partir d'une sélection de textes, d'un module de visualisation des thématiques extraites ainsi que du réseau social sous-jacent. La plate-forme est utile tant pour la veille de presse (en permettant l'identification rapide des sujets importants dans les forums) que pour la recherche sur médias sociaux (en permettant aux chercheurs de constituer des corpus dynamiques de données textuelles).

Références

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören. Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Web Semantics : Science, Services and Agents on the World Wide Web*, 7(3) :154–165, 2009.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3 :993–1022, 2003.

[3] Claudiu Musat, Julien Velcin, Stefan Trausan-Matu, and Marian-Andrei Rizoiu. Improving topic evaluation using conceptual knowledge. In *International Joint Conference on Artificial Intelligence, Proceedings of the Twenty-Second*, volume 3 of *IJCAI 2011*, pages 1866–1871. AAAI Press, 2011.

[4] Marian-Andrei Rizoiu and Julien Velcin. Topic extraction for ontology learning. In Wilson Wong, Wei Liu, and Mohammed Bennamoun, editors, *Ontology Learning and Knowledge Discovery Using the Web : Challenges and Recent Advances*, chapter 3, pages 38–61. Hershey, PA : Information Science Reference, 2011.

[5] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Structuring typical evolutions using temporal-driven constrained clustering. In *International Conference on Tools with Artificial Intelligence, Proceedings of the Twenty-Forth*, ICTAI 2012, pages 610–617. IEEE, November 2012.

[6] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Unsupervised feature construction for improving data representation and semantics. *Journal of Intelligent Information Systems*, 2013.

[7] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *International Conference on Machine Learning, Proceedings of the Eighteenth*, pages 577–584. Morgan Kaufmann, 2001.

[8] Djamel A. Zighed, Shusaku Tsumoto, Zbigniew W. Ras, and Hakim Hacid, editors. *Mining Complex Data*, volume 165 of *Studies in Computational Intelligence*. Springer, 2009.