La fouille de données

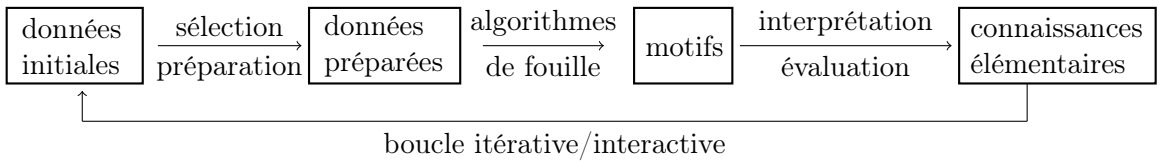
Amedeo Napoli, Alexandre Termier

Quelques notes sur la fouille de données

Amedeo Napoli[1] et Alexandre Termier[2]
[1] LORIA (CNRS - Inria - Université de Lorraine)
BP 239, 54506 Vandoeuvre les Nancy (Amedeo.Napoli@loria.fr)
[2] Centre de recherche INRIA/IRISA
Campus de Beaulieu, 35042 Rennes (Alexandre.Termier@irisa.fr)

1 Introduction

À l'heure actuelle, des données — les documents sur le web par exemple — sont disponibles en quantité importante et en qualité variable, sans finalité particulière et précise a priori. Plusieurs questions fondamentales se posent : (i) est-il possible de comprendre ce que renferment ces masses de données, (ii) est-il possible d'y découvrir et d'en extraire "quelque chose d'intéressant", (iii) comment s'y prendre de façon efficace pour mener à bien de telles opérations. Les éléments de réponse vont servir à résoudre des problèmes de recherche d'information, de classification, de recommandation, et d'ingénierie des connaissances (représentation et raisonnement) C'est là que le *processus de découverte de connaissances dans les (bases de) données* (DCD) montre tout son intérêt, qui peut se définir comme la recherche dans de grands volumes de données de motifs présentant suffisamment d'intérêt pour être réutilisés. L'expression "motif" est à prendre au sens large et s'assimile à une règle, une classe d'individus, une description ensembliste, une séquence, un arbre, un graphe ... La DCD cherche à passer de données brutes (signaux) à des informations (données interprétées) puis à des connaissances (informations prêtes à l'emploi). Ce processus est typique de l'apprentissage machine, avec des enjeux comme le passage à l'échelle pour le temps (efficacité des algorithmes) et l'espace (volumes à traiter), l'interprétation et la réutilisation des motifs en termes de connaissances à destination des agents humains et logiciels.



Classiquement, le processus de DCD se divise en trois étapes principales : (1) la préparation des données, (2) la fouille des données, et (3) l'interprétation des motifs extraits. De plus, le processus est itératif et interactif : un "analyste" contrôle le processus et l'oriente selon ses objectifs. In fine, les motifs extraits peuvent être interprétés, en fonction d'un modèle du domaine des données, comme des éléments de connaissances de ce domaine et à ce titre réutilisables dans un système de connaissances.

Les méthodes de fouille de données sont numériques ou symboliques. Les premières relèvent essentiellement des statistiques, des probabilités et de l'analyse des données. Les secondes s'appuient sur les arbres de décision, la recherche de motifs et de règles d'association, l'analyse formelle de concepts, la classification ("clustering"), ... La DCD fait donc appel à

une variété de techniques propres à l'apprentissage, à la gestion des bases de données, à l'ingénierie des connaissances, à la calculabilité et plus généralement aux mathématiques discrètes et du continu.

2 La recherche de motifs

Considérons l'exemple célèbre dit du "panier de la ménagère" où les données sont les tickets de caisse des clients d'un ou de plusieurs supermarchés. Un ticket de caisse est constitué d'un ensemble d'articles achetés par un client. Un *motif* correspond ici à un ensemble d'articles achetés dans un même ticket, par exemple "*Pain, Beurre, Chocolat*". Un motif a un *support*, le nombre de tickets qui contiennent le motif, et une *fréquence* qui est le support rapporté au nombre total de tickets. Si notre motif "*Pain, Beurre, Chocolat*" a une fréquence de 17%, cela signifie qu'il est présent dans 17% des tickets de caisse, ce qui peut s'avérer une connaissance commerciale importante : elle peut par exemple aider à organiser physiquement les rayons du magasin en éloignant au maximum ces trois articles afin d'allonger le parcours du client. À partir d'un motif on peut extraire des règles, par exemple "$Beurre, Chocolat \rightarrow Pain$". Cette règle a une *confiance*, qui est la probabilité conditionnelle qu'un client ayant acheté du *Beurre* et du *Chocolat* achète aussi du *Pain*. Là encore cette règle fournit une connaissance importante : si la règle a une confiance élevée, disons 80%, cela signifie que faire une promotion sur le beurre et/ou le chocolat augmentera mécaniquement les ventes de pain.

Étant donné une base de données éventuellement très volumineuse — des millions de tickets et des milliers d'articles par exemple —, un problème majeur consiste à énumérer l'ensemble des motifs présentant certaines caractéristiques, comme par exemple être fréquents. C'est un problème calculatoire difficile, où la difficulté ne peut pas être simplement contournée en rajoutant des machines. En effet, si A est l'ensemble des articles référencés dans le supermarché, la recherche de motifs consiste à parcourir un espace d'états qui correspond à l'ensemble des parties de A dont la taille est $2^{|A|}$ où $|A|$ est le cardinal de A (le nombre d'articles). Dans un supermarché standard il y a au minimum des dizaines de milliers de références ($|A| > 10000$), donc le nombre de motifs dépasse les 2^{10000}, un chiffre beaucoup plus grand que le nombre estimé de particules dans l'univers ($\sim 10^{85}$) !

Toutefois, en général seule une infime partie de ces motifs vérifient les caractéristiques voulues comme la fréquence. Les chercheurs en fouille de données ont pu exploiter cette observation et certaines propriétés mathématiques de l'espace de recherche pour proposer des algorithmes efficaces d'énumération de motifs, dont le plus ancien représentant est l'algorithme Apriori (1994).

La recherche de motifs a été étendue à des données et motifs plus complexes, ayant des structures de séquences, d'arbres ou de graphes. Un exemple simple d'extraction de motifs dans une séquence est l'analyse d'ADN : la donnée d'entrée est une séquence d'ADN et les motifs sont toutes les sous-séquences se répétant fréquemment dans cette séquence. On peut chercher les motifs séquentiels de manière stricte — la sous-séquence ATG a été trouvée 324 fois — ou souple — la sous-séquence $ATG{**}GC$ a été trouvée 128 fois, où '*' peut être n'importe quel caractère. Les motifs de type graphes eux matérialisent des sous-graphes se répétant dans une base de données de graphes. Ils sont particulièrement utiles dans l'analyse de molécules et de réactions chimiques. Ainsi, les motifs extraits correspondent à des sous-structures moléculaires spécifiques pouvant servir de support à des médicaments.

La recherche de ces types de motifs plus complexes s'appuie sur le même principe qu'indiqué plus haut : il y a un espace de recherche gigantesque à explorer, mais certaines propriétés permettent d'élaguer des parties importantes de cet espace. Découvrir ces propriétés et les ex-

ploiter au mieux est nécessaire pour avoir des algorithmes de découverte de motifs capables de passer à l'échelle sur des volumes réalistes de données. Toutefois, à l'heure actuelle, le challenge principal en découverte de motifs est ailleurs. En effet, l'exécution des méthodes classiques d'énumération de motifs peut produire des millions de motifs potentiellement intéressants. Un analyste humain ne peut pas exploiter une telle quantité d'information. Les travaux actuels se concentrent donc sur la mise au point de nouvelles méthodes découvrant un petit nombre de motifs particulièrement intéressants. Ainsi, après avoir étudié et développé de nombreux algorithmes très efficaces (aspect quantitatif), la communauté en fouille de données s'oriente actuellement vers des aspects plus qualitatifs liés aux connaissances du domaine étudié faisant intervenir des mesures d'intérêt, des contraintes ou encore des préférences.

3 Références

Les conférences les plus importantes du domaine de la fouille de données sont ECML-PKDD ("European Conference on Machine Learning and Principles and Practice of Knowledge Discovedy"), KDD ("Conference on Knowledge Discovery and Data Mining"), ICDM ("International Conference on Data Mining") et SDM ("SIAM Data Mining"). Les journaux principaux sont DMKD ("Data Mining and Knowledge Discovery"), Machine Learning, TKDD ("Transactions on Knowledge Discovery from Data") et TKDE ("Transactions on Knowledge and Data Engineering"). De nombreuses références existent sur le sujet et parmi les ouvrages de base complets et détaillés, nous voudrions citer :

1. Charu C. Aggarwal. Data Mining : The Textbook. Springer 2015.

2. Jiawei Han, Micheline Kamber, Jian Pei : Data Mining : Concepts and Techniques (Third Edition). Morgan Kaufmann, 2011.

3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar : Introduction to Data Mining. Pearson, 2005.

4. Mohammed J. Zaki and Wagner Meira Jr. Data Mining and Analysis – Fundamental Concepts and Algorithms, Cambridge University Press, 2014.