Post-Doc : Co-clustering sur données de très grande taille

ref : 0014645 | 02 Nov 2016
apply before : 30 Jan 2017
2 avenue Pierre Marzin 22300 LANNION - France

Pour postuler : https://orange.jobs/jobs/offer.do?do=fiche&id=57600

about the role

Objectifs scientifiques - résultats attendus

Khiops coclustering est un outil d'analyse exploratoire qui permet l'analyse de la corrélation entre deux ou plusieurs variables catégorielles ou numériques, basé sur une approche de sélection de modèle appelé MODL (Boullé, 2006, 2001). Cet outil (disponible sur www.khiops.com) est de plus en plus utilisé, avec des applications dans une variété de problèmes :

- Marketing : les clients avec la liste des produits achetés (customer x product).
- Web Mining : analyse de logs web pour identifier des comportements de navigation (cookies x webpages).
- Télécommunications : dimensionnement de réseau mobile par l'analyse des compte-rendus d'appels (CDRs) (sourceAntenna x targetAntenna), e.g. analyse exploratoire des CDRs à l'échelle d'un pays (Guigourès, 2013).
- Fouille de textes : (co)clustering de textes (Texts x Words).
- Fouille de graphes : données multigraphes temporels (SourceNodes x TargetNodes xTime), e.g., analyse des locations de vélos à Londres (Guigourès et al., 2012).
- Clustering de données fonctionnelles (séries temporelles numériques ou catégorielles) : TimeSeriesId x Time xValue ou TimeSeriesId xTime x Event, e.g. clustering de courbes (Boullé, 2012).

Khiops coclustering est capable de traiter des données d'assez grande taille, avec des millions d'instances et des dizaines de milliers de valeurs par variable catégorielle, avec une complexité algorithmique sous-quadratique par rapport au nombre d'instances. L'outil atteint par contre ses limites pour le traitement de données de très grande taille, comportant potentiellement des milliards d'instances pour des variables ayant des millions de valeurs. C'est par exemple le cas si l'on souhaite analyser les CDR à l'échelle d'un pays, en passant d'une granularité au niveau des antennes (applications en dimensionnement de réseau) à une granularité au niveau des clients (application en marketing avec identification de communautés extrêmement fines et gestion individualisée de l'expérience utilisateur). Le graphe à résumer est alors de taille trop importante pour les algorithmes actuels de co-clustering.

L'objectif du post-doc est d'étendre les algorithmes d'optimisation de co-clustering aux graphes de très grande taille, pour le critère MODL (Boullé, 2011) de co-clustering. Parmi les

pistes envisagées, on pourrait utiliser ou adapter des algorithmes de partitionnement hiérarchique de graphe de type H-metis (Karypis ert al, 2000 ; Selvakkumaran et al, 2006), adaptés aux graphes de très grande taille. On pourrait alors obtenir un partitionnement initial « grossier », permettant ensuite d'appliquer l'algorithme de co-clustering classique à des sous-parties du graphe. L'ensemble des co-clusterings partiels peut alors être réconcilié dans une dernière passe pour produire un co-clustering global. Ce type algorithme en trois passes (co-clustering initial, co-clusterings partiels, réconciliation) peut être généralisé à plusieurs niveaux de hiérarchie et permettre une parallélisation des algorithmes utilisés. Parmi les alternatives à l'utilisation de H-metis pour la première passe, on peut citer l'utilisation d'algorithmes de Singular Value Decomposition (SVD) pour le clustering de graphes de grande taille, ou plus simplement l'application d'un co-clustering initial sur un échantillon de petite taille, permettant ensuite de projeter le reste des données sur les co-clusters obtenus, avant application des passes suivantes.

about you

Vous avez déjà effectué une thèse dans le domaine des statistiques et des mathématiques.

Une expérience est souhaitée dans le domaine statistique et dans le développement informatique.
Des connaissances en apprentissage statistique sont un réel plus.

Des compétences en programmation sont nécessaires: au minimum, une excellente maîtrise d'un langage de script dédié à l'analyse de données (R, Matlab, Python avec bibliothèque Scikit-learn…).

Une forte motivation, des capacités de synthèse, de rédaction (article de conférence ou de revue) et de présentation des travaux (anglais), et à s'intégrer dans une équipe sont également demandées.

additional information

Références

[1] Bock, H. : Simultaneous clustering of objects and variables. Analyse des données et Informatique, 1979.

[2] M. Boullé. MODL: a Bayes optimal discretization method for continuous attributes. Machine Learning, 65(1):131-165, 2006.

[3] M. Boullé. Data grid models for preparation and modeling in supervised learning. In Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1, I. Guyon, G. Cawley, G. Dror, A. Saffari (eds.), pp. 99-130, Microtome Publishing, 2011.

[3] M. Boullé. Functional data clustering via piecewise constant nonparametric density estimation. Pattern Recognition, 45(12):4389-4401, 2012.

[5] Cheng, Y. et Church, G. M. (2000). : Biclustering of expression data. In Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB), 2000.

[6] Dhillon, I. S., Mallela, S., et Modha, D. S. : Information-theoretic co-clustering. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 89{98. ACM.2003.

[7] Govaert, G. et Nadif, M. : Clustering with block mixture models. Pattern

Recognition, 2003.

[8] Govaert, G. and Nadif, M. : Co-Clustering. ISTE Ltd and John Wiley & Sons Inc, 2013.

[9] R. Guigourès, M. Boullé, F. Rossi. A Triclustering Approach for Time Evolving Graphs. In Co-clustering and Applications, IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012), Pages 115-122, 2012

[10] R. Guigourès. Utilisation des modèles de co-clustering pour l'analyse exploratoire. Phd thesis. 2013.

[11] Hartigan, J. A. : Clustering Algorithms. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

[12] G. Karypis and V. Kumar. Multilevel k-way Hypergraph Partitioning. VLSI Design, Vol. 11, No. 3, pp. 285 - 300, 2000.

[13] N. Selvakkumaran and G. Karypis. Multi-Objective Hypergraph Partitioning Algorithms for Cut and Maximum Subdomain Degree Minimization. IEEE Transactions on CAD, 25(3), pp. 504-517, 2006.

[14] Seung, D. et Lee, L. : Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13, pages 556-562, 2001

[15] Xu, G, Zong, Y, Dolog, P. et Zhang, Y. : Co-clustering Analysis of Weblogs Using Bipartite Spectral Projection Approach. Knowledge-Based and Intelligent Information and Engineering Systems, 2010

[16] Yoo, J. et Choi, S. : Orthogonal nonnegative matrix tri-factorization for co-clustering : Multiplicative updates on stiefel manifolds. Information processing & management, 2010.

department

Vous serez dans l'équipe de traitement statistique de l'information d'Orange Labs Lannion. Cette équipe spécialisée en machine learning, data mining et profiling, comporte une vingtaine de permanents, sur des sujets allant de la recherche aux applications opérationnelles, ainsi qu'une demi-douzaine de doctorants et post-doc.

Vous pouvez trouver ici des exemples de publications scientifiques de l'équipe, pour 2015, illustrant ces thématiques :

http://vincentlemaire-labs.fr/publis/PublicationsScientifiquesPROF_2015.pdf

Contexte global de l'étude et état de l'art (bibliographie)

Avec la disponibilité massive de données (big data, open data, linked data), la valorisation des données devient un enjeu critique pour Orange ainsi que tous ses clients. Orange dispose de nombreuses données dans plusieurs contextes applicatifs, et il est souvent nécessaire de commencer toute étude par une analyse exploratoire. Par exemple, dans le cas de CRM (Customer Relationship Management), les données textuelles de type verbatim sont potentiellement pertinentes pour la prédiction de scores de churn ou d'appétence, mais leur représentation initiale n'est pas adaptée à la modélisation directe. Il est alors nécessaire de procéder à une analyse exploratoire préalablement à l'utilisation de ces données dans des modèles de scoring. Dans le cas de données d'usage d'un nouveau service, il est intéressant d'explorer les données de façon à mettre en évidence une segmentation utilisateurs permettant par la suite d'adapter les actions marketing.

Face à ces besoins, la maîtrise de techniques d'analyse exploratoire efficaces est un enjeu majeur pour être plus performant que les autres opérateurs de données.

Proposée initialement par (Hartigan, 1975), le co-clustering est une extension du clustering simple (standard). Le principe du clustering simple est de former des groupes d'individus similaires entre eux et différents des individus appartenant à d'autres groupes. L'avantage du co-clustering réside dans l'étude simultanée (jointe) entre deux types d'entités qui permet d'extraire la structure sous-jacente existante entre elles. Les méthodes de co-clustering sont donc particulièrement adaptées à l'analyse exploratoire dans le cas de données de liens entre plusieurs entités en relation, disponibles dans de nombreux contextes applicatifs : text mining (texte x mots), basket analysis (customer x product), web log mining (user x page), analyse de détails de communication (caller x called)…

Plusieurs méthodes ont été développées pour extraire des structures sous-jacentes dans les données à l'aide de méthodes de co-clustering (Bock, 1979; Cheng et al, 2000; Dhillon et al, 2003; Xu et al, 2010). Ces méthodes diffèrent principalement selon le type des données traitées, les hypothèses considérées, la méthode d'extraction utilisée et les résultats souhaités. En particulier, il existe plusieurs grandes familles de méthodes pour effectuer la classification croisée :

- Méthodes de reconstruction de matrices qui réécrit le problème de classification sous forme d'approximation matricielle (Seung et al, 2001 ; Yoo et al, 2010, Xu et al, 2010), CROEUC, CROBIN, CROKI2 (pour les données continues, binaires et de contingence (Govaert, 1983)).
- Méthodes basées sur les modèles probabilistes : utiliser des variables latentes dans un modèle de mélanges pour définir les blocs (Govaert et al, 2003; Govaert et al, 2013).
- Les méthodes de co-clustering basées sur l'approche MODL (Boullé, 2011) exploitent des modèles probabilistes pour deux à plusieurs entités de type quelconque (numérique ou catégoriel), ne nécessitent aucun paramètre et bénéficient d'algorithmes de complexité sous-quadratique permettant de traiter des données de grande taille.

contract

Post Doc