❑ **COLLABORATION**

INSTITUT DE MATHEMATIQUES DE TOULOUSE

❑ **DOCUMENTS RELATIFS AUX TRAVAUX**

Léa Laporte, Rémi Flamary, Stéphane Canu, Sébastien Déjean et Josiane Mothe. Non-convex Regularizations for Feature Selection in Ranking with Sparse SVM. IEEE Transactions on Neural Networks and Learning Systems, Vol. 25 N. 6, p. 1118-1130, 2014. https://www.irit.fr/publis/SIG/2014_OTAO_LLFRCSDSMJ.pdf -

Léa Laporte, Sébastien Déjean et Josiane Mothe. *Séparateurs à Vaste Marge pondérés en norme L2 pour la sélection de variables en apprentissage d'ordonnancement.* Conférence en Recherche d'Information et Applications (CORIA), 19/03/2014 – 21/03/2014, Nancy.

Léa Laporte. *La sélection de variables en apprentissage d'ordonnancement pour la recherche d'information : vers une approche contextuelle*. Rapport de thèse, 2013.

❑ **PARTICIPANTS**

LEA LAPORTE

SEBASTIEN DEJEAN

JOSIANE MOTHE

❑ **CONTACTS**

J. MOTHE
Institut de Recherche en Informatique de Toulouse
Equipe Systèmes d'Informations Généralisées
Campus Université Paul Sabatier
118, Route de Narbonne – 31062 Toulouse Cedex 4
Email : josiane.mothe@irit.fr,
Tél. : (33/0) 561 55 64 44 Fax.: (33/0) 561 55 62 58

EQUIPE **SYSTEMES D'INFORMATIONS GENERALISEES – J. MOTHE**

FEATURE SELECTION IN RANKING WITH WITH SPARSE REWEIGHTED L2-NORM SVM.

RANKRWFS AND RANKL2-AROM ALGORITHMS

☐ Skill ☑ Prototype ☐ Product

UMR 5505 CNRS-INP-UPS 118 ROUTE DE NARBONNE 31062 TOULOUSE CEDEX 4

❑ FEATURE SELECTION FOR RANKING IN INFORMATION RETRIEVAL

En Recherche d'Information (RI), l'apprentissage d'ordonnancement vise à optimiser de façon automatique l'ordonnancement des documents restitués par les systèmes de recherche d'information. Les méthodes apprennent des fonctions d'ordonnancement qui combinent les résultats de modèles de recherche d'information pour classer les documents. Les scores des modèles de RI sont ainsi les caractéristiques des fonctions apprises. Le nombre de caractéristiques utilisées posent des problèmes de volumétrie, pouvant entraîner une augmentation importante des temps d'exécution des algorithmes d'apprentissage et des temps de réponse des systèmes en ligne. L'incorporation de méthodes de sélection de variables aux approches d'ordonnancement permet de résoudre ce problème.

❑ REWEIGHTED L2-NORM SUPPORT VECTOR MACHINES FOR FEATURE SELECTION IN RANKING

En apprentissage et plus particulièrement dans le cadre des SVM ou de méthodes linéaires, une façon naturelle d'effectuer la sélection de variables est de considérer un problème régularisé avec une norme L0 (nombre d'éléments non nuls dans le vecteur des poids des caractéristiques). Or, la régularisation L0 est non convexe, non différentiable et non continue : le problème d'optimisation à résoudre est NP-complexe. En général, les travaux de la littérature préfère considérer la norme L1 (somme des valeurs absolues des poids du vecteur) qui est convexe. Néanmoins, elle est non différentiable, les algorithmes d'optimisation quadratique ne peuvent être utilisés. Les approches de type Moindres Carrés Pondérés (Iteratively Reweighted Least Squares - IRLS) permettent d'approcher les problèmes d'optimisation en norme L0 et L1 par itérations successives de problèmes en norme L2. Nous avons proposé deux types d'algorithmes basés sur des adaptations de méthodes IRLS pour la classification : RankL2-AROM et RankRWFS.

❑ RANKL2-AROM AND RANKRWFS ALGORITHMS

L'algorithme RankL2-AROM est une adaptation de l'approche L2-AROM. Elle approche le problème régularisé en norme L0 par itérations successives de problèmes en norme L2. A chaque itération, pour chaque observation, la valeur de chaque caractéristique est multipliée par le coefficient trouvé pour cette caractéristique. Les caractéristiques dont le coefficient est inférieur à un seuil donné sont supprimés du modèle. L'algorithme s'arrête lorsque le nombre de caractéristiques restantes souhaité est atteint (fourni par l'utilisateur).
L'algorithme RankRWFS permet d'approcher un problème en norme L0 ou en norme L1. La pondération des observations est spécifique à la régularisation considérée. Les caractéristiques dont le coefficient est inférieur à un seuil donné sont supprimés du modèle. L'algorithme s'arrête lorsque le nombre de caractéristiques restantes souhaité est atteint (fournit pas l'utilisateur).

MODULES

– *Sélection, apprentissage et production des fichiers de résultats* : un script permet de spécifier la méthode à utiliser, le nombre de caractéristiques souhaitées, le jeu de données à utiliser. Des fichiers contenant les modèles finaux d'une part, les valeurs de précision moyenne, NDCG et MAP d'autre part, sont produits.

– *Repondération de la norme L2* : un script implémentant les boucles pour RankL2-AROM et un script implémentant les boucles pour RankRWFS sont disponibles.

DEVELOPPEMENTS

- L'application a été développée en Matlab, mais peut être exécutée indifféremment sur Matlab ou Octave (libre d'accès, gratuit). Le solveur des SVM en norme L2 utilisé est RankSVM-Primal, implémenté en Matlab par Olivier Chapelle (Utilisation à des fins de recherche uniquement, disponible en ligne à l'adresse suivante : http://olivier.chapelle.cc/primal/ranksvm.m).

APPLICATIONS

- *Sélection de variables et ordonnancement sur les jeux de données issues des collections LETOR*