Extension de l'algorithme CURE aux fouilles de données volumineuses

Jerzy Korczak et Aurélie Bertaux

LSIIT, Bd. Sébastien Brant, 67412 Illkirch cedex France

<korczak, bertaux>@lsiit.u-strasbg.fr

Dans ce poster, nous allons proposer une démarche pour découvrir le fonctionnement du cerveau en se basant sur un concept de fouille de données. Ce concept peut se définir comme l'extraction de connaissances potentiellement exploitables à partir d'images IRMf. C'est une approche interactive qui intègre directement l'expert-médecin dans le processus de découverte et d'apprentissage de concepts pour mettre en évidence les zones fonctionnelles du cerveau et leur organisation.

CURE selon Guha et al. (1998) est un algorithme de classification, mais il est robuste face aux outliers et permet d'identifier des groupes non sphériques et d'une grande variance de taille. CURE réalise ceci en représentant chaque groupe par un nombre fixé de points qui sont générés en sélectionnant des points bien dispersés du groupe, et ensuite rapprochés du point moyen au centre du groupe en le multipliant par un coefficient. Le fait d'avoir plus d'un point représentatif permet à CURE de bien s'ajuster à la géométrie des clusters non sphériques et l'opération de rapprochement de ses points permet de diminuer les effets des outliers.

Pour manipuler de grandes volumes de données, CURE emploie une combinaison d'échantillonnage aléatoire et de partitionnement. Un échantillon tiré de l'ensemble des données et tout d'abord partitionné et chaque partition est partiellement mise en cluster. Chacun de ces groupes partiels sera à nouveau regroupé lors d'une seconde passe de l'algorithme pour extraire les clusters désirés.

Une force de CURE, selon les auteurs, est de pouvoir s'adapter à de grandes bases de données pour un algorithme hiérarchique. L'implémentation de la version originale a démontré certaines faiblesses de performances de la classification de signaux tels que ceux de l'IRMf est très lourde car il s'agit de voxels à laquelle s'ajoute la quatrième dimension de leur évolution dans le temps. Pour réduire le temps de classification, nous avons proposé quelques améliorations.

Tirage aléatoire. Un tirage aléatoire des données est utilisé ayant pour vertu d'améliorer la qualité de la classification car les signaux sont enregistrés selon l'ordre dans lequel l'IRM les balayent, ce qui fait que deux signaux qui sont issus de zones voisines peuvent être séparés lors de l'enregistrement. En effet, toute une couche est balayée dans un sens avant de passer à la couche inférieure.

Echantillonnage. Cela permet de déterminer les classes, avec moins de signaux. Ce cas est extrêment important car CURE fonctionnant de manière hiérarchique plus le nombre de signaux est important, plus il génère de classes et plus les calculs entre toutes les classes prennent du temps et des ressources.

Partitionnement. Sur cette même constation, un système de rechargement en signaux a été réalisé. CURE classant les clusters par ordre croissant de leur distance au cluster qui leur est le plus proche, impose donc un calcul de distance entre chaque paire de clusters, et pour chaque paire, leur distance est la distance minimale entre toutes les paires des signaux représentatifs des deux classes. Nous avons déterminé expérimentalement un nombre fixe maximum de clusters à traiter ensemble. Pas à pas l'algorithme fusionne deux à deux les clusters jusqu'à atteindre un seuil fixé à partir duquel nous effectuons un rechargement en nouveaux clusters pour réatteindre le nombre maximal fixé. Ce procédé est répété jusqu'à épuisement du nombre de signaux.

La plateforme d'expérimentation de fouille d'images IRMf a été développée par Korczak et al. (2005) comprenant des algorithmes de classification de signaux IRMf et permettant une fouille visuelle interactive en temps quasi réel. Plusieurs algorithmes ont déjà été implémentés notamment : K-means, LGB, SOM et GNG.

L'algorithme CURE a été testé sur des données simples bi-dimmensionnelles et sur des données synthétiques et comparés aux autres algorithmes déjà implémentés suivant les protocoles décrits par Hommet (2005). Les classifications ont été réalisées par variation respective des paramètres que sont le nombre de classes, le rapport de dilution des voxels activés et le rapport signal sur bruit. Si sur les données simples, CURE obtient une très bonne performance cependant, il s'avère que sur les données synthétiques, il présente des performances moyennes, mais reste de bonne robustesse. Cette constatation ne concerne que des données synthétiques ne lui permettant pas de mettre en avant ses qualités d'adaptation à des clusters d'une morphologie non sphérique.

En tant qu'algorithme hiérarchique, CURE est extrêment gourmand en ressources. Nos améliorations ont réduit la complexité algorithmique et en conséquence ont réduit les temps de calculs. Selon la simulation on peut envisager une utilisation d'algorithme CURE étendue avec des contraintes de temps réel.

Références

Guha, R. Rastogi, K. Shim (1998). *CURE : An Efficient Clustering Algorithm for Large Databases*. SIGMOD 1998, pages 73-84.

Hommet, J (2005). *Fouille interactive de séquences d'images 3D d'IRMf*. Rapport de LSIIT, CNRS, Illkirch.

Korczak, J., C. Scheiber, J. Hommet, N. Lachiche (2005). *Fouille interactive en temps réel de séquences d'images IRMf.* Numéro Spécial RNTI.

Summary

In this poster, an extended unsupervised data mining algorithm CURE is briefly described and evaluated. CURE is used to extract active voxels from brain images and is compared with several other unsupervised algorithms on fMRI images.