

# 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 extrêmement 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.