Clustering hiérarchique basée sur une approche multi-agents

Lieu du stage :
- Laboratoire d'Informatique de Grenoble (LIG)
- Equipe Analyse de données, Modélisation et Apprentissage automatique (AMA)

Responsables :
- Gilles Bisson (LIG/AMA), gilles.bisson@imag.fr
- Catherine Garbay (LIG/AMA), catherine.garbay@imag.fr

Contexte
L'apprentissage non-supervisé (ou clustering) est un processus exploratoire visant à regrouper les instances d'une base de données en un ensemble de classes *contrastées* et *homogènes*. Il correspond à l'activité bien connue en Scinece Cogntive de "Catégorisation". Par exemple, partant d'une base contenant les clients d'une entreprise, le processus mettra en évidence des types de consommateurs. Dans ce contexte, la classification ascendante hiérarchique (CAH) organise, de surcroît, ces classes sous la forme d'un arbre binaire strict dont les nœuds sont les concepts appris et les feuilles les instance ; l'information apprise par l'algorithme est ainsi structurée selon différents niveaux de généralités : les classes les plus générales étant à la racine de l'arbre. Cette méthode est largement utilisée dans tous les domaines d'activités techniques et scientifiques.

Toutefois pour certaines catégories de problème, construire un arbre binaire n'est pas sémantiquement optimum. Par exemple, lorsque l'on applique la CAH pour classer les mots d'une collection de documents afin de construire une ontologie (dictionnaire structuré de termes), on constate rapidement que la structure obtenue :
- contient beaucoup trop de niveaux de généralités intermédiaires du fait que l'on élabore un arbre binaire, typiquement en langage naturel le nombre de niveaux de généralité attendu ne dépasse pas la dizaine
- ne permet pas de mettre en évidence la polysémie des mots. Par exemple, un mot comme "saumon" pourrait-être rattaché soit à un concept de "couleur", soit à un concept de "animal" ce qui est impossible avec un arbre stricte.

Dans un tel contexte, il serait beaucoup plus naturel que la structure classificatoire utilise un graphe acyclique. Or, cette direction n'a été que très peu étudiée dans la littérature en apprentissage et souvent sur des problématiques un peu différentes se rattachant notamment à la classification à partir de données décentralisées (Parunak et al. 06), (Reed et al. 04).

Déroulement du stage
Dans la CAH la construction des classes repose sur un opérateur unique d'agrégations des classes deux à deux en partant des instances. L'objectif de cours de ce stage est donc de concevoir et développer un algorithme ou plusieurs *agents de classifications* différents (fusion, partage, …), mis en concurrence, permettront de construire une structure plus complexe qu'un arbre binaire. L'idée étant d'avoir une méthode plus proche de ce que fait un être humain lorsqu'il doit classer un ensemble d'informations.

Dans un premier temps, il s'agira pour le stagiaire d'explorer la littérature pour voir quelles sont les approches qui ont été explorées. Ensuite, il devra définir de manière formelle les agents à développer en relation avec les objectifs de la classification, puis concevoir une architecture logicielle (inspirée de la CAH) intégrant les critères d'application et de contrôle des agents. On s'appliquera à mettre en évidence les propriétés (convergence, …) vérifiées par cette architecture. L'algorithme résultant sera à implémenter de préférence en Python, le domaine d'application privilégié étant ici celui la construction automatique d'ontologie à partir de collection de textes. Enfin, durant le stage, en fonction du temps disponible, d'autres aspects fondamentaux pourront être explorés par le candidat, notamment:
- La parallélisation de la méthode afin de diminuer la complexité en temps et/ou mémoire
- La conception d'agents plus "cognitifs" permettant d'obtenir une méthode de classification incrémentale capable de faire évoluer dynamiquement la structure hiérarchique en

fonction de l'ajout de nouvelles instances. Idéalement, la classification pourrait même combiner des approches ascendante et descendante.

Prérequis

Le stage se situe clairement à la frontière entre les sciences cognitives et l'informatique. Le candidat ne devra pas être rebuté par la nécessité d'implémenter un logiciel, ni de faire des évaluations de performances.

Gilles Bisson | Email : gilles.bisson@imag.fr
Charge de recherche CNRS Equipe AMA - Laboratoire LIG - UMR 5217 | Fax :

Centre Equation 4 - UFR IM2AG - BP 53 - F-38041 Grenoble Cedex 9
