Clustering hiérarchique multi-opérateur

Responsables :

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

Site de l'équipe AMA : http://ama.liglab.fr

Contexte de la recherche

L'apprentissage non-supervisé (ou clustering) peut être vu comme la « mécanisation » du processus de catégorisation chez les êtres vivants. En informatique, il s'agit de regrouper les instances d'une base de données en un ensemble de classes *contrastées* et *homogènes*. 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.

Toutefois, construire un arbre binaire n'est pas toujours sémantiquement optimum. Par exemple, lorsque l'on applique la CAH pour classer une collection de documents textuels soit pour organiser ces documents en catégories thématiques, soit de manière duale pour construire une ontologie (dictionnaire structuré de termes), on constate que le structure obtenue :

- contient beaucoup trop de niveaux de généralités intermédiaires du fait que l'on élabore un arbre binaire,
- ne permet pas de mettre en évidence la polysémie des documents ou des mots. Par exemple, un documents peut appartenir à différentes catégories et de même un mot tel "saumon" peut-être rattaché soit à un concept de "couleur", soit à un concept de "animal" ce qui est impossible avec une hiérarchie 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 instances/classes deux à deux en utilisant une distance prédéfinie. Ici on s'appuiera sur la notion de co-similarité (Bisson et al.12) qui permet de comparer des textes utilisant des termes différents. 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.

Dans un premier temps, il s'agira pour le (ou la) stagiaire d'explorer la littérature pour recenser qu'elles sont les approches similaires 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 de la catégorisation automatique de collection de textes (ou de termes) en collaboration avec la société Short-Edition qui est un éditeur communautaire d'histoires courtes.

Durant le stage, en fonction du temps disponible, d'autres aspects fondamentaux pourront être explorés par le stagiaire, notamment :

- La parallélisation de la méthode afin de diminuer la complexité en temps et/ou mémoire
- La conception d'agents 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

Une expérience raisonnable de la programmation est nécessaire (python ou autre langage).