Programmation séquentielle - Classification

1 Buts

— Implémenter un algorithme de classification.
— Manipuler les notions de liste et d'arbre binaire avec des pointeurs.

2 Enoncé

Partant d'une liste de nombres saisis au clavier, on construit une liste simplement chaînée triée. Par transformations successives, on obtient un arbre binaire où chaque nœud contient la moyenne de ses 2 noeuds enfants. Ceci est réalisé en groupant à chaque étape les nombres les plus proches dans la liste. A la fin, les feuilles de l'arbre contiennent les nombres de départ (voir un exemple à la fig. 1).

2.1 Cahier des charges

1. Définir un nœud capable de contenir un nombre et deux pointeurs.
2. Définir une liste simplement chaînée dont les éléments sont précisément ces nœuds.
3. Définir un arbre dont les éléments sont les nœuds définis ci-dessus.
4. Écrire une fonction permettant d'insérer des nombres réels saisis au clavier dans une liste en ordre trié. Ces nombres sont immédiatement placés dans un nœud avec les sous-arbres gauche et droit initialisés à null.
5. Écrire une fonction qui construit l'arbre de classification par transformation de la liste ci-dessus. A chaque étape, la moyenne entre les deux éléments les plus proches de la liste est calculée, elle remplacera l'un des deux éléments, l'autre sera supprimé. A chaque étape, la liste contiendra un élément de moins et ceci jusqu'à ce qu'il n'en reste plus qu'un.
6. Écrire une fonction qui affiche l'état de la liste chaînée et l'arbre sous-jacent à chaque élément. La liste s'affiche verticalement et l'arbre correspondant à la liste s'affiche horizontalement (voir le cours).

3 Mise en contexte

L'exploration de données, aussi connue sous le nom de fouille de données (data mining en anglais), a pour objet l'extraction de connaissances à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. La classification en est un sous-domaine qui vise à regrouper des données en fonction de leur proximité.

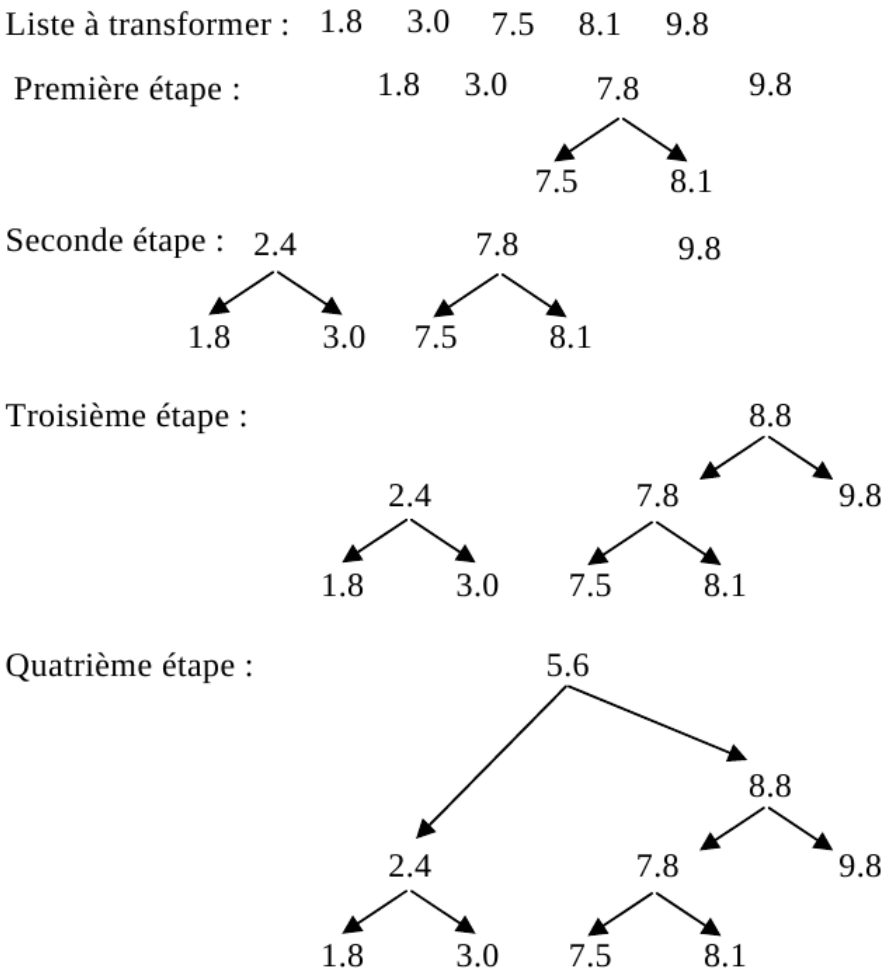

FIGURE 1 – Un exemple de transformation de liste en arbre.

La classification hiérarchique ascendante entend organiser sous forme d'arbre binaire un ensemble de données dont on peut mesurer la proximité via une notion de distance. On procède par regroupements successifs de paires de groupes de données. Initialement, chaque donnée forme un groupe (singleton). A la première étape sont regroupées les deux données les plus proches. Puis, à chaque étape, sont fusionnés les deux groupes dont la distance est la plus faible. Le processus s'arrête quand les deux groupes restant fusionnent en un unique groupe contenant toutes les données. La question de la définition de la distance entre deux groupes de données est centrale.

La phylogénie est l'étude de la formation et de l'évolution des organismes vivants en vue d'établir leur parenté. La classification phylogénétique est un système de classification des êtres vivants basé sur la classification hiérarchique ascendante. Elle utilise la notion de proximité évolutive des espèces pour construire des dendrogrammes comme illustré sur la fig. 2.

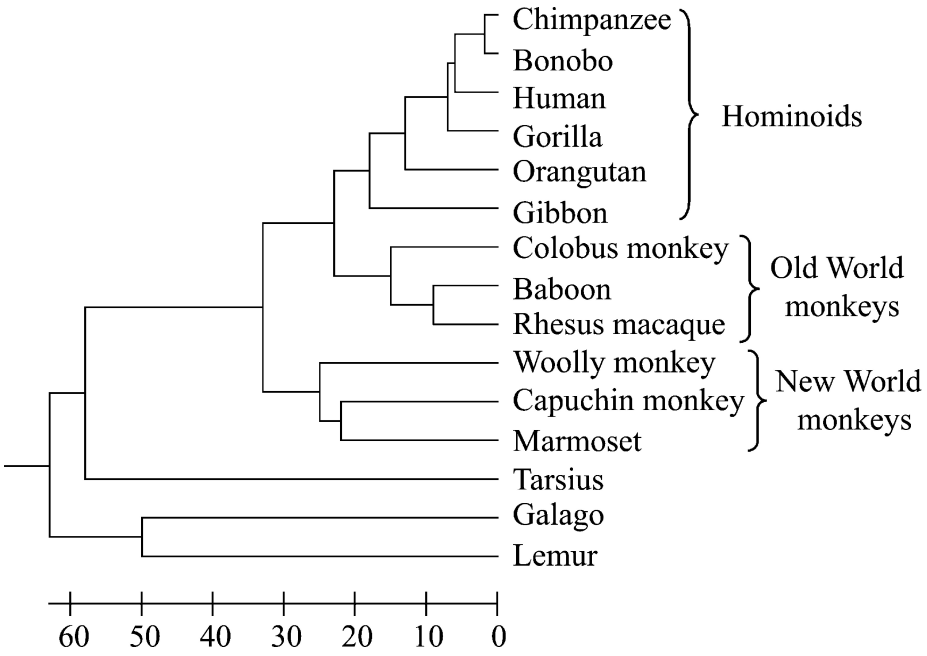

FIGURE 2 – Un exemple d'arbre philogénétique.