# "Un outil pour la classification à base de clustering pour décrire et prédire simultanément "

Vincent Lemaire, Oumaima Alaoui Ismaili

Cette démonstration présente un logiciel (accessible à tous) intégrant un algorithme de "k-moyennes supervisées" produisant un modèle (« déployable ») et des rapports décrivant les clusters obtenus.

Depuis quelques années, les chercheurs ont concentré leur attention sur l'étude d'un nouveau aspect d'apprentissage connu sous le nom de la classification à base de clustering (ou Supervised clustering en anglais) (e.g., (Eick et al., 2004) et (Cevikalp et al., 2007)). Les approches appartenant à ce type d'apprentissage visent à décrire et à prédire d'une manière simultanée (Alaoui Ismaili et al., 2015a). Dans ce cadre d'étude, on suppose que la classification à base de clustering est étroitement liée à l'estimation de la distribution des données conditionnellement à une variable cible. A partir d'une base de données étiquetée, ces approches cherchent à découvrir la structure interne de la variable cible afin de pouvoir prédire ultérieurement la classe des nouvelles instances.

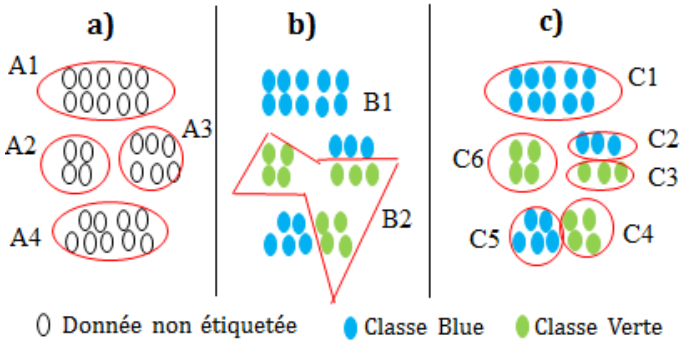


FIG. 1: Types d'apprentissage

La figure 1 illustre la différence entre les trois types d'apprentissage : le clustering standard a), la classification supervisée b) et la classification à base de clustering c). Dans la classification supervisée, la compacité des classes apprises (dans la phase d'apprentissage) n'est pas une condition importante (e.g, le groupe B2 de b)). Le clustering regroupe les instances homogènes sans tenir en compte leur étiquetage (e.g, le groupe A4 de la a)). La classification à base de clustering vise à former des groupes compacts et purs en termes de classes (e.g, les 6 groupes de c)).

La classification à base de clustering est très utile dans les domaines critiques où l'interprétation des résultats fournis par un système d'apprentissage est une condition primordiale. Elle permet à l'utilisateur de découvrir les différentes voies qui peuvent mener à une même prédiction : par exemple de découvrir que deux instances de même classe peuvent être très hétérogènes (e.g., les instances appartenant au groupe C1 et au groupe C5 de c)). La classification à base des K-moyennes est une version modifiée de l'algorithme des K-moyennes standard. Elle cherche à générer des partitions ayant un bon compromis entre la compacité des groupes formés et leurs puretés en termes de classes (voir la partie c) de la figure ci-dessus). La prédiction de la classe des nouvelles instances se réalise par la suite en se basant sur la structure interne découverte lors de la phase d'apprentissage.

Eick, C. F., N. Zeidat, et Z. Zhao (2004). Supervised clustering-algorithms and benefits. In Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, pp. 774–776. IEEE.

Cevikalp, H., D. Larlus, et F. Jurie (2007). A supervised clustering algorithm for the initialization of rbf neural network classifiers. In Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th, pp. 1–4. IEEE.

Alaoui Ismaili, O., V. Lemaire, et A. Cornuéjols (2015a). Classification à base de clustering ou comment décrire et prédire simultanément. In Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA).

---

Note : Ce logiciel se nomme

 Khiops Ennéade

Il est protégé par l'enregistrement FR.001.520021.000.S.P.2012.000.00000 to the Agency of Software Protection.

Il peut être testé gratuitement pour une période d'évaluation.

Il est diffusé commercialement (en dehors d'Orange) par la société **Predicsis**