Clustering dynamique d'un flot de données : un algorithme incrémental et optimal de détection des maxima de densité

Alain Lelu

LASELDI / Université de Franche-Comté
30 rue Mégevand – 25030 Besançon cedex
alain.lelu@univ-fcomte.fr

Résumé. L'extraction non supervisée et incrémentale de classes sur un flot de données (*data stream clustering*) est un domaine en pleine expansion. La plupart des approches visent l'efficacité informatique. La nôtre, bien que se prêtant à un passage à l'échelle en mode distribué, relève d'une problématique *qualitative*, applicable en particulier au domaine de la veille informationnelle : faire apparaître les évolutions fines, les « signaux faibles », à partir des thématiques extraites d'un flot de documents. Notre méthode GERMEN localise de façon exhaustive les maxima du paysage de densité des données à l'instant t, en identifiant les perturbations locales du paysage à t-1 et modifications de frontières induites par le document présenté. Son caractère optimal provient de son exhaustivité (à une valeur du paramètre de localité correspond un ensemble unique de maxima, et un découpage unique des classes qui la rend indépendante de tout paramètre d'initialisation et de l'ordre des données.

1 Introduction et objectifs

Pour rendre compte avec exactitude des évolutions temporelles, cruciales dans beaucoup de domaines d'application (ex. : veille d'information), il est nécessaire à notre avis :
1) de partir d'une base stable, c'est-à-dire d'une classification :
- indépendante de l'ordre de présentation des données (exigence n°1),
- indépendante des conditions initiales, que ce soit d'un choix de « graines de classes » arbitraires ou dépendantes des données (exigence n°2),
- impliquant un minimum de paramètres, un seul si possible, pour réduire l'espace des choix et tendre vers un maximum de vérifiabilité et de reproductibilité (exigence n°3).
2) d'ajouter aux contraintes d'une bonne classification celle de l'incrémentalité (exigence N°4), afin de saisir les évolutions au fil de l'eau : rectifications de frontières entre classes, apparition de nouvelles classes, voire de « signaux faibles »... Pour nous, il y a incrémentalité véritable si le résultat de la classification est indépendant de l'ordre des données présentées antérieurement (exigence N°5), tout en découlant des données antérieures, par un historique pouvant faire l'objet d'interprétations.

Notre démarche a été de concevoir une méthode où la contrainte d'incrémentalité participer d'un tout cohérent, en vue d'aboutir à tout instant à une classification qui ait du sens, et dont la différence de représentation par rapport à l'instant précédent ne provient que des