SuSE: Subspace Selection embedded in an EM algorithm

Laurent Candillier*,**, Isabelle Tellier*, Fabien Torre*, Olivier Bousquet**

*GRAppA - Université Charles de Gaulle - Lille 3
**Pertinence - 32 rue des Jeûneurs - 75002 Paris

Le *subspace clustering* [Parsons et al. (2004)] est une extension du *clustering* traditionnel [Berkhin (2002)] qui recherche un ensemble de *clusters* qui peuvent être définis dans différents sous-espaces. C'est le cas, par exemple, des données présentées dans la figure 1. L'intérêt de telles techniques est important dans le cadre de données contenant un nombre important de dimensions car elles permettent de faire face à la *malédiction de la dimensionalité*. De plus, elles permettent de fournir une description réduite des clusters obtenus car les clusters sont alors définis par un nombre restreint de dimensions. Or, la problématique de la compréhensibilité des résultats obtenus en clustering est également importante.

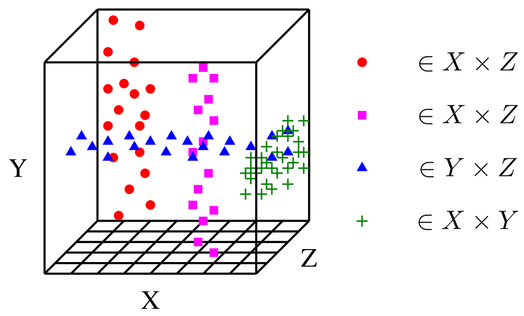

FIG. 1 – *Exemple de 4 clusters définis dans différents sous-espaces.*

Dans l'étude que nous présentons, nous mettons en avant l'intérêt d'utiliser des modèles probabilistes pour le subspace clustering. En particulier, nous montrons que le problème difficile de la spécification des paramètres des méthodes de subspace clustering peut être vu comme un problème de *sélection de modèle* dans un cadre probabiliste. Ceci nous permet alors de développer un algorithme de subspace clustering qui, contrairement aux méthodes existantes, ne nécessite aucune connaissance a priori de la part de l'utilisateur.

Nous mettons également en avant l'intérêt de permettre le chevauchement des clusters dans ce cadre. De plus, le problème de la détection du bruit qui peut exister dans les données s'intègre naturellement dans une telle modélisation probabiliste. Et nous montrons qu'être capable de détecter le bruit permet de fournir un résultat plus compréhensible. Enfin, la prise en compte de différents types de dimensions est également facilitée dans le cadre probabiliste.

Nous avons donc développé et testé un tel algorithme de subspace clustering statistique nommé **SuSE**, ainsi qu'une méthode originale permettant de fournir une représentation compréhensible des clusters obtenus. Différentes expérimentations, menées sur des données aussi bien artificielles que réelles, ont mis en avant la pertinence de notre approche dans le cadre de la détection des clusters et de leurs sous-espaces spécifiques, ainsi que dans celui de la compréhensibilité des résultats produits.

SuSE est basé sur le modèle classique de mélange de distributions de probabilités et l'utilisation de l'algorithme EM [Ye et Spetsakis (2003)]. Un paramètre noté k spécifie le nombre de clusters du modèle, et un autre paramètre noté d le nombre de dimensions utilisées pour caractériser ces clusters. La méthode globale de **SuSE** est alors de générer différents modèles avec différentes valeurs pour ces paramètres k et d, puis de sélectionner le modèle le plus approprié aux données.

La figure 2 montre l'intérêt du critère BIC [Ye et Spetsakis (2003)], basé sur le calcul de la vraisemblance du modèle par rapport aux données et la pénalisation des modèles plus complexes, dans ce cadre de sélection de modèle. Elle montre, sur données artificielles, l'évolution de la valeur du critère BIC en fonction de la différence entre le nombre réel K de clusters et le nombre k de clusters recherchés, et de la différence entre le nombre réel D et le nombre sélectionné d de dimensions utiles pour caractériser ces clusters. On remarque dès lors que la valeur optimale du critère BIC est atteinte lorsque le nombre de clusters recherchés et le nombre sélectionné de dimensions utiles atteignent les nombres réels.

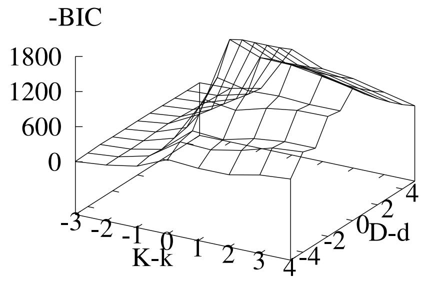

FIG. 2 – *Évolution de la valeur du critère BIC en fonction de* $(K - k)$ *et de* $(D - d)$ *[pour une meilleure visualisation, -BIC est reporté au lieu de BIC].*

Références

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, California.

Parsons, L., E. Haque, et H. Liu (2004). Evaluating subspace clustering algorithms. In *Workshop on Clustering High Dimensional Data and its Applications, SIAM Int. Conf. on Data Mining*, pp. 48–56.

Ye, L. et M. Spetsakis (2003). Clustering on unobserved data using mixture of gaussians. Technical report, York University, Toronto, Canada.