1 Introduction

Lorsque nous appliquons un algorithme de classification non supervisée (clustering) à un ensemble d'objets décrits par un ensemble de variables, la partition obtenue dépend de plusieurs paramètres dont le critère optimisé qui dans la plupart des cas s'appuie explicitement ou implicitement sur une mesure de dissimilarité. D'autre par, la partition obtenue dépend de la méthode choisie. Par conséquent, si nous avons deux algorithmes de clustering différents et que nous les appliquons au même jeu de données, nous pouvons obtenir des résultats très différents. Mais quel est le bon? Comment pouvons-nous évaluer les résultats? Contrairement aux méthodes de classification supervisée, l'objectif est plus compliqué. Néanmoins l'approche *consenus clustering* peut être utilisée dans le cadre non supervisé; elle s'avère très intéressante.

L'idée de combiner différents résultats de clustering (*cluster ensemble* ou *clustering aggregation*) est apparue comme une approche alternative pour améliorer la qualité des résultats des algorithmes de classification même avec des nombres de classes différents. Étant donné un ensemble d'objets, une méthode d'ensemble clustering comprend deux étapes principales: la génération, qui consiste à créer un ensemble de partitions de ces objets (différentes initialisations, différents algorithmes, différentes modèles etc.), et la fonction Consensus, où une nouvelle partition, qui est un *compromis* entre toutes les partitions obtenues à l'étape de génération, est calculée.

Pusieurs approches *ensemble* existent et la litérature est abondante. Vous disposez de papiers, du code dans plusieurs packages disponibles; je vous recommande ce papier. Pour évaluer le *consensus clustering*, des données de différents types sont également disponibles.

2 Sélection de modèle pour un nombre de classes connu

Disposant d'un ensemble d'objets décrits par des variables continues, voir Table 1. Il est possible d'envisager de faire du consensus clustering de la manière suivante.

Table 1: Description des données images: nombre de lignes, colonnes et classes

datasets	# samples	# features	# classes
JAFFE	213	676	10
MNIST5	3495	784	10
MFEA	2000	240	10
USPS	9298	256	10
OPTIDIGITS	5620	64	10

1. Considérer tous les modèles Gaussiens avec un nombre de classes égal à 10.
2. Obtenir à l'aide de EM la meilleure partition, voir le package `mclust`.
3. A partir de cet ensemble de partitions appliquer un EM dérivant d'un modèle multinomial général, voir le package `Rmixmod`; en effet il s'agit d'un tableau de modalités. La partition obtenue peut être vue éventuellement comme une partition (avec le même nombre de classes) *consensuelle*. Est-elle meilleure que la meilleure partition obtenue avec le modèle choisi par le critère BIC, en termes de Accuracy, NMI et ARI ?
4. En considérant un nombre restreint de partitions, il est possible d'améliorer les résultats en considérant éventuellement les meilleures partitions au lieu de toutes les partitions. Dans ce cas, envisager des stratégies pour améliorer vos résultats, justifier votre démarche.

3 Information Mutuelle

L'objectif est d'améliorer le clustering de l'ensemble des documents mais en s'appuyant cette fois-ci sur un algorithme de co-clustering. Nous considérons ici l'algorithme basé sur l'information mutuelle, voir `CoClust`.

Nous envisageons d'évaluer les performances de cet algorithme sur les quatre bases dans données choisies dans ce papier à télécharger.

1. Lancer `CoClustInfo` plusieurs fois, retenir les 5-10 meilleures partitions (au sens du critère optimisé). Prendre le même nombre de classes en lignes et en colonnes (le vrai nombre de classes).

2. A partir de ces 5-10 partitions, proposer un consensus clustering en utilisant les méthodes s'appuyant sur les hypergraphes `CSPA`, `HGPA` et `MCLA` dans `Cluster_Ensembles`.

3. En utilisant l'approche basée sur les matrices de co-associations. Dans une matrice de *co-association*, chaque cellule de la matrice est égale si deux objets appartiennent à la même classe et 0 sinon, construire une matrice de *co-association* totale qui est la somme des 5-10 matrices de *co-association* obtenues dans question 1. Sur cette matrice appliquer `CoClustInfo`, que peut-on dire ?.

4. Reprendre la question 1. mais en ajoutant la partition obtenue à l'aide de `Spherical k-means` ? Que peut-on dire ?

5. Refaire l'étude en augmentant le nombre de classes en colonnes (2x, 3x).

4 Modularité

La modularité est communément utilisée dans les graphes et notamment dans les réseaux sociaux. Elle peut être également utilisée pour réaliser du co-clustering (avec le même nombre de classes en ligne et en colonnes). Elle apparait même comme un outil efficace pour détecter le nombre de co-clusters. Dans `CoClust` vous disposez de deux algorithmes `CoClustSpecMod` et `CoClustMod`.

1. Reprendre les mêmes questions 1 à 4 dans Section 3 pour chacun des deux algorithmes `CoClustSpecMod` et `CoClustMod`.

2. Que donne un consensus en combinant les partitions des lignes de `CoClustInfo`, `CoClustSpecMod`, `CoClustMod` et `Spherical k-means` ?

5 Modèle de von Mises-Fisher (*facultatif mais...*)

Les modèles de vMF sont connus pour leur intérêt pour considérer les données directionnelles comme c'est la cas des données document-terme. En s'appuyant sur les packages `movMF` et `skmeans`, et 4 bases au choix disponibles dans le package `CLUTO` dédié au clustering des matrices document-terme, voici les tâches à réaliser.

1. Reprendre les questions de Section 2, mais en considérant cette fois-ci le modèle vMF au lieu du modèle Gaussien et les 3 algorithmes considérés

2. Vous avez noté que le critère optimisé par `Spherical k-means` est associé à un modèle vMF sous contraintes. Faites varier le nombre de classes et retenir la meilleure partition. Que donne le consensus sur les meilleures partitions avec différents nombres de classes.

Recommandation

Le rapport doit se limiter à 6 pages double-colonne (maximum), le format demandé vous sera envoyé, le code commenté doit être disponible en dehors du rapport. Dans ce rapport,

1. il est nécessaire de décrire brièvement l'(les) approche(s) consensus considéré(es),

2. les tables de données étudiées et les tables de comparaison entre les différentes algorithmes, partitions et approches doivent être présentes,

3. les commentaires des résultats sont indispensables pour chaque chaque question,

4. le rapport peut être rédigé en Français ou en Anglais,
5. enfin, ce travail est à réaliser par binôme ou seul, il doit comprendre une conclusion et une bibliographie.

N.B. Une grande partie de l'examen en Janvier portera essentiellement sur ce projet.