Thème 5 : méthodes avancées

Partie 2: Forage de données de hautes dimensions : Projected Clustering

shengrui.wang@usherbrooke.ca

Plan

1- Motivations

2- Projected Clustering

3- PROCLUS*

4- PCKA*

Pour cette partie il suffit de bien saisir les motivations et les problématiques associés au « projected clustering », c'est-à-dire les sections 1 et 2. Les deux algos des sections 3 et 4 sont fournis comme exemples seulement et ne font pas l'objet de l'examen final.

Motivations (constats des algorithmes vues jusqu'à maintenant)

- Le processus de clustering vise à construire des groupes(clusters) d'objets similaires à partir d'un ensemble hétérogène d'objets.
- Le clustering repose sur une mesure précise de la similarité des objets que l'on veut regrouper. Cette mesure est appelée distance ou métrique.
- La mesure de distance la plus utilisée est la distance euclidienne.
- Pour mesurer la distance entre deux points, on doit prendre en considération de toutes les dimensions.
- Deux objets sont considérés comme similaires s'ils ont des valeurs très proches dans toutes les dimensions.

Motivations

❑ Problème de la distance avec les données de hautes dimensions

- Dans l'espace de grande dimension, il est fort probable que, pour n'importe quels pairs de points qui appartiennent au même cluster, il existe seulement peu de dimensions dont les points sont proches les uns des autres.

➢ Le concept de la similarité devient invalide; car:

- Une fonction de distance accorde la même importance à toutes les dimensions. Cependant, toutes les dimensions ne sont pas au même titre d'égalité

Motivations

❑ Problème de la distance avec les données de hautes dimensions

- Des études* théoriques récentes démontrent que :

➢ La différence relative des distances entre différents points décroit avec l'augmentation du nombre de dimensions.

➢ Plus que la dimension de l'ensemble de données augmente plus la distance entre les points devient la même!

* http://citeseer.ist.psu.edu/cache/papers/cs/11824/http:zSzzSzwww.cs.wisc.eduzSr~jgoldstzSrnnpaper.pdf/beyer99when.pdf

Motivations

- Exemple 1

Person	Age	Virus Level	Blood Type	Disease A
1	35	0.95	AB	Uninfected
2	64	0.9	AB	Uninfected
3	27	1.0	AB	Uninfected
4	18	9.8	O	Infected
5	42	8.6	AB	Infected
6	53	11.3	B	Infected
7	37	0.75	O	Recovered
8	28	0.8	A	Recovered
9	65	0.89	B	Recovered

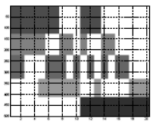
Motivations

Exemple 2

➢ On collecte l'évaluation d'un grand nombre de clients sur certain nombre de produits.

▪ Dans notre ensemble de données les lignes représentent les clients et les colonnes (les dimensions) représentent les produits.

▪ Le but est d'identifier les clients qui partagent les mêmes préférences.



➢ Les clusters des clients existent dans différents sous-espace de dimensions.

Motivations

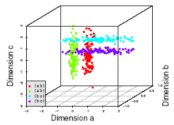
Exemple 3

	A_1	A_2	A_3	A_4	A_5
x_1	0	0	0	10	100
x_2	0	0	0	70	30
x_3	20	10	20	50	50
x_4	80	80	20	50	50

Si nous utilisons la distance euclidienne comme mesure de similarité, il est fort probable que x_2 et x_3 vont être placés dans le même cluster, puisque leur distance (c.-à-d. 41.23) est la plus petite comparativement aux distances entre n'importe quels pairs d'objets. Cependant, une simple inspection visuelle suggère qu'il y a deux clusters : $C_1 = \{x_1, x_2\}$ et $C_2 = \{x_3, x_4\}$. Les dimensions $\{A_1, A_2, A_3\}$ forment le cluster C_1, alors que les dimensions $\{A_3, A_4, A_5\}$ forment le cluster C_2. Les sous-ensembles de dimensions $\{A_4, A_5\}$ et $\{A_1, A_2\}$ représentent les dimensions non pertinente pour C_1 et pour C_2 respectivement.

Motivations

Exemple 4



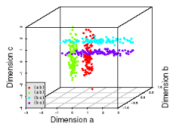
Ensemble de données avec 4 clusters.

➢ *cluster 1 et cluster 2 existent dans les dimensions a & b.*

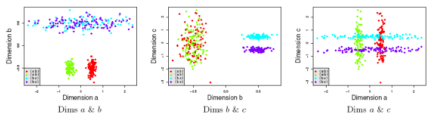
➢ *cluster 3 et cluster 4 existent dans les dimensions b & c.*

Motivations

Exemple 4



Projection sur deux dimensions



Motivations

Les données de grandes dimensions sont caractérisées par

➢ Les clusters existent dans différents sous-espace de dimensions (c.à.d. les clusters n'existent pas forcément dans les mêmes dimensions).

➢ Par conséquent

✓ Une dimension peut être complètement inutile (ne contient aucune une structure de clusters ➔ aucune région dense)

✓ Une dimension peut contenir juste une partie de l'information utile (un mélange de régions denses et bruits)

✓ Une dimension peut contenir toute l'information utile (la dimension en question contient que des régions dense, aucun bruit)

Motivations

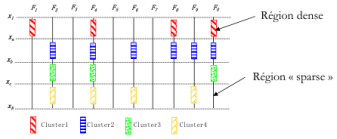
Est-ce que les techniques de réduction de dimensions peuvent apporter une solution à ce problème ?

➢ **Non, car:**

Les techniques de réduction de dimensions adoptent une approche globale : Une dimension est sélectionnée en entier ou rejeté en entier alors que les clusters les clusters peuvent exister dans différentes combinaisons des sous-espaces de dimensions et non dans tout l'ensemble des dimensions. En d'autres termes: certaines dimensions peuvent être discriminantes pour la formation d'un certain cluster, alors que ces mêmes dimensions peuvent s'avérer peu pertinente pour la formation d'un autre cluster.

Motivations

Exemple 5



Cluster 1 existe dans F1, F4, F8, F10
Cluster 2 existe dans F2, F4, F6, F8, F9, F10
Cluster 3 existe dans F2, F4, F10
Cluster 4 existe dans F2, F4, F6, F9

Motivations

Exemple 5

- Aucun algorithme de clustering* n'est capable d'identifier les 4 clusters.

- Les techniques de sélection/réduction de dimensions peuvent réduire la dimension de cet ensemble de données et ce en éliminant F3, F5 et F7, mains il y a un grand risque qu'ils éliminent aussi l'attribut F1 car il contient beaucoup de bruit alors que cluster 1 existe dans F1.

➔ L'utilisation des techniques de sélection/réduction de la dimension peu causer une perte d'information.

* Les algorithmes qui utilisent une fonction de distance qui donne la même importance à toutes les dimensions

Plan

Projected Clustering

Cluster projeté : « projected cluster »

Un sous-ensemble S des points, associé à un sous-ensemble des dimensions D, tels que les points de S soient fortement groupés (proche l'un de l'autre) dans le sous-espace des dimensions D.

Cluster 1 : (S1, D1) = ({x1, x2}, {F1, F2, F3})
Cluster 2 : (S2, D2) = ({x3, x4}, {F3, F4, F5})

	F1	F2	F3	F4	F5
x1	0.1	2.12	0.9	40	100
x2	0.15	2.10	0.95	70	80
x3	100	20	6	50	0.5
x4	10	80	6	49.6	0.6

- Les résultats d'un algorithme de Projected Clustering :
  - ✓ La répartition des points dans des clusters
  - ✓ Les dimensions associées à chaque cluster

Projected Clustering : un algorithme simple, une extension de k-moyenne

Nous pouvons étendre k-means pour identifier les dimensions importantes pour chaque cluster. Cette extension se réalise par l'introduction des poids propres à chaque cluster et par l'intégration d'un processus d'optimisation dans l'algorithme.

Pour chaque cluster k et chaque dimension (ou variable) j, créer une variable de poids w_{kj} .

Le nombre total de poids w_{kj} est : le nombre-de-clusters * nombre-de-attributs

w_{kj} >=0 et la norme des poids pour chaque cluster $\|w_{k\cdot}\| = C$, i.e. la norme reste constante. C pourrait être 1

Projected Clustering : un algorithme simple

Modification de la mesure de distance : la distance entre un objet x et le centre d'un cluster k, i.e. v_k est redéfinie comme

$$d(x, v_k, w_{k,\cdot}) = \sqrt{\sum_{i=1}^{D} w_{k,i}(x_i - v_{k,i})^2}$$

Ici D est le nombre des dimensions

Projected Clustering : un algorithme simple

Mise-à-jour de l'ensemble des poids $\{w_{kj}\}$: La mise-à-jour se fera après une époque d'itérations de clustering couvrant tous les objets. Elle se réalise en deux étapes :

1) Décrémenter w_{kj} pour pénaliser les dimensions de grande variance. Pour chaque cluster k et chaque dimension j,

$$w_{k,j} = \frac{w_{k,j}}{1 + variance \;\; de \;\; variable \;\; j \;\; dans \;\; clusterk}$$

Ici, la variance de la variable j dans chaque cluster k est calculée individuellement. Plus que la variance est grande, plus w_{kj} est réduite relativement.

Projected Clustering : un algorithme simple

Mise-à-jour de l'ensemble des poids $\{w_{kj}\}$: La mise-à-jour se fera après une époque d'itérations de clustering couvrant tous les objets. Elle se réalise en deux étapes :

2) La deuxième étape de la mise-à-jour consiste à la normalisation comme suit,

$$w_{k,j}^{new} = \frac{C * w_{k,j}}{\sqrt{\sum_{i=1}^{D} (w_{k,i})^2}}$$

Cette formule permet de rétablir $\|w_{k.}\| = C$. Le résultat final de l'intégration de cette mise-à-jour de deux étapes dans k-means serait la diminution graduelle des poids associés aux dimensions dans lesquelles la forme de cluster est allongée.

Évaluation

Comment évaluer les dimensions identifiées de chaque cluster ?

Il existe 3 indices :

$$DR_1(C_s) = \frac{\sum_{x_i \in C_s, F_j \in F_s} (x_{ij} - x_{sj})^2 / d_S}{\sum_{x_i \in C_s, F_j \in F} (x_{ij} - x_{sj})^2 / d}$$

$$DR_2(C_s) = \frac{\sum_{x_i \in C_s, F_j \notin F_s} (x_{ij} - x_{sj})^2 / (d - d_S)}{\sum_{x_i \in C_s, F_j \in F} (x_{ij} - x_{sj})^2 / d}$$

$$DR_3(C_s) = \frac{\sum_{x_i \notin C_s, F_j \in F_s} (x_{ij} - x_{sj})^2 / d_S}{\sum_{x_i \notin C_s, F_j \in F} (x_{ij} - x_{sj})^2 / d}$$

Pour un bon clustering
DR1 < 1
DR2 > 1
DR3 > DR1