5

K-Prototype

5. K-Prototype

K-Means est l'un des algorithmes de clustering les plus (sinon le plus) utilisés, ce qui n'est pas surprenant. Il est rapide, a une implémentation robuste dans sklearn, et est intuitivement facile à comprendre.

K-Prototypes est un frère moins connu mais qui offre l'avantage de travailler avec des types de données mixtes.

Il mesure la distance entre les caractéristiques numériques en utilisant la distance euclidienne (comme K-means) mais mesure également la distance entre les caractéristiques catégorielles en utilisant le nombre de catégories correspondantes.

Il a été publié pour la première fois par Huang (1998) et a été implémenté en python en utilisant ce package.

K-Prototype est une méthode de clustering basée sur le partitionnement.

Son algorithme est une forme améliorée de l'algorithme de clustering K-Means et K-Mode pour gérer le clustering avec des types de données mixtes.

Source

5. K-Prototype

Voici les étapes simples de **l'algorithme** K-Prototype :

1. Sélectionnez k prototypes initiaux dans l'ensemble de données X. Il doit y en avoir un pour chaque cluster.
2. Attribuer chaque objet dans X à un cluster dont le prototype est le plus proche de lui. Cette attribution est effectuée en tenant compte de la mesure de dissimilarité décrite ci-après.
3. Une fois que tous les objets ont été attribués à un cluster, testez à nouveau la similarité des objets par rapport aux prototypes actuels. Si vous constatez qu'un objet est trouvé tel qu'il est le plus proche d'un autre prototype de cluster, mettez à jour les prototypes des deux clusters.
4. Répétez l'étape 3, jusqu'à ce qu'aucun objet ne change de cluster après avoir entièrement testé X.

Comment les objets sont-ils répartis dans les clusters ?

$$E_l = \sum^{n} y_{il} \sum^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum^{n} y_{il} \sum^{m_c} \delta(x_{ij}^c, q_{lj}^c)$$

C'est en considérant la mesure de dissimilarité.
Mesure de dissimilarité
Soit X={x1, x2,..., xn} un ensemble de n objets.
Xi = [xi1, xi2,..., xim] chaque objet est représenté par m valeurs d'attributs.
Les objets de X sont d'abord divisés en k clusters disjoints.
Quel est ce prototype ? C'est le centre du cluster.

$$E_l = E_l^r + E_l^c$$

1-1er groupe
r - Attributs numériques
c - Attributs catégoriels
yi1 - L'objet Xi appartient au cluster 1
r1 - Poids de l'attribut catégorique dans le cluster 1
Si r1 est petit, cela indique que le clustering est dominé par les attributs numériques.
Si r1 est grand, cela indique que le clustering est dominé par des attributs catégoriels.
E1= Somme minimale de la différence entre tous les éléments et les prototypes de la grappe 1
E1r= Somme minimale de la différence des attributs numériques de tous les éléments et des prototypes de la grappe 1
E1c= Somme minimale de la différence des attributs catégoriels de tous les éléments et des prototypes de la grappe 1

xi1 - jème attribut
qij - jème attribut du prototype dans le cluster 1
mr - Nombre d'attributs numériques
mc - Nombre d'attributs catégoriels
m = mr + mc

5. K-Prototype

Huang [6] a proposé un algorithme k-prototypes pour le regroupement de données de type mixte, qui combine les idées de l'algorithme k-means et de l'algorithme k-modes. L'algorithme k-prototypes divise l'ensemble de données en différents sous-clusters $k(k \in N^{+})$ afin de minimiser la valeur de la fonction de coût. La fonction de coût est représentée par la formule suivante :

$$F(U,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} u_{il} d(x_i, q_l).$$

L'algorithme k-prototypes combine les "moyennes" de la partie numérique et les "modes" de la partie catégorielle pour construire un nouveau "prototype" de centre de regroupement hybride. Sur la base du "prototype", il construit une formule de coefficient de dissimilarité et une fonction de coût applicables aux données de type mixte. Le paramètre gamma est introduit pour contrôler l'influence de la caractéristique catégorielle et de la caractéristique numérique sur le processus de clustering. On suppose que l'ensemble de données de type mixte possède une caractéristique numérique et une caractéristique catégorielle. Pour n'importe quel , la définition du coefficient de dissimilarité de k-prototypes est présentée dans la formule suivante :

$$d(x_i, q_l) = \gamma \sum_{s=1}^{p} \delta\left(x_{i,s}^{C} - q_{l,s}^{C}\right) + \sum_{s=p+1}^{m} \sqrt{\left(x_{i,s}^{N} - q_{l,s}^{N}\right)^2},$$

$$\text{where} \quad \delta(x_{i,s}, q_{l,s}) = \begin{cases} 0, x_{i,s} = q_{l,s}, \\ 1, x_{i,s} \neq q_{l,s}. \end{cases}$$

5. K-Prototype

L'algorithme des k-prototypes divise le coefficient de dissimilarité des données de type mixte en deux parties pour un calcul séparé.
La partie catégorielle adopte la distance de Hamming simple, et la partie numérique adopte le carré de la distance euclidienne.
La proportion des deux types de données dans le coefficient de dissimilarité a été ajustée par le paramètre gamma. Il s'agit d'un paramètre réglable important pour l'algorithme des k-prototypes.
Le but de l'introduction du paramètre est d'éviter la déviation de la valeur du résultat du clustering à partir de la caractéristique catégorielle ou de la caractéristique numérique et de contrôler le poids relatif de la dissimilarité entre les données catégorielles et les données numériques.
Lorsque $x_{i,s}^{C} = q_{l,s}^{C}$, $\delta(x_{i,s}^{C}, q_{l,s}^{C})$, est égal à 0 ; lorsque $x_{i,s}^{C} \neq q_{l,s}^{C}$, $\delta(x_{i,s}^{C}, q_{l,s}^{C})$ est égal à 1 ; les étapes de base de l'algorithme des k-prototypes sont décrites comme suit :

Etape 1 : les objets de données ont été sélectionnés au hasard dans l'ensemble de données comme centres de clusters initiaux.
Étape 2 : la formule (2) est utilisée pour calculer la dissimilarité entre x1 et q1. Selon le résultat du calcul, xi est attribué au cluster le plus proche.
Étape 3 : selon les centres de cluster actuels, la dissimilarité de l'objet de données est recalculée. Réaffecter les objets de données au sous-cluster le plus proche, les valeurs avec la fréquence la plus élevée sont utilisées dans la partie catégorielle, et la partie numérique utilise la méthode de la valeur moyenne pour déterminer. Mettez à jour les centres des clusters.
Étape 4 : répétez les étapes 2 et 3 jusqu'à ce que la fonction de coût ne change plus. Si la fonction de coût ne change plus, l'algorithme se termine. Sinon, passez à l'étape 2 pour continuer.

5. K-Prototype

Avantages :

L'algorithme des k-prototypes permet de regrouper des données de type mixte.
Le principe est simple et facile à mettre en œuvre

Inconvénients :

(1) La sélection aléatoire des centres de clusters initiaux entraîne l'incertitude et le caractère aléatoire des résultats de clustering, et le nombre de clusters k doit être déterminé manuellement.
(2) la simple distance de Hamming est utilisée pour calculer la dissimilarité entre les données catégorielles et les centres de clusters, ce qui entraîne une perte d'informations et l'incapacité de refléter objectivement la situation réelle entre les objets de données et les clusters, d'où des résultats de clustering inexacts.
(3) le paramètre gamma utilisé pour ajuster la proportion entre les données catégorielles et les données numériques doit être déterminé manuellement.
(4) les caractéristiques structurelles des données catégorielles et des données numériques et la distribution globale des ensembles de données n'ont pas été pleinement prises en compte.

5. K-Prototype