Apport de l'entropie pour les c-moyennes floues sur des données catégorielles

Abdoul Jalil Djiberou Mahamadou*, Violaine Antoine*
Engelbert Mephu Nguifo*, Sylvain Moreno**

*Université Clermont Auvergne, CNRS, LIMOS, ENSMSE
LIMOS, F-63000 Clermont Ferrand France
{abdoul_jalil.djiberou_mahamadou, violaine.antoine, engelbert.mephu_nguifo}@uca.fr,
**Digital Health Hub, Université Simon Fraser, Vancouver, Canada
sylvain_moreno@sfu.ca

1 Introduction

La méthode de clustering flou des *c-moyennes* avec centroids flous FC (Kim et al., 2004) est une extension de la méthode fuzzy k-modes (FKM) (Huang et Ng, 1999) utilisant une représentation floue des centres des clusters. Après dérivation de la fonction objectif de cette méthode, il a été démontré dans (Djiberou Mahamadou et al., 2020) que la formule de mise à jour des centres ne permet pas de garantir la convergence de la méthode. Par la suite, les auteurs ont proposé deux extensions de FC dénommées FC* et CFE (Categorical Fuzzy Entropy c-means). Tandis que FC* utilise des mises à jour dures des centres, CFE incorpore la notion d'entropie dans la fonction objectif pour jouer un rôle de pénalisation sur les poids. Cela permet ainsi une répartition de la masse des poids plus équilibrée sur toutes les valeurs de l'attribut considéré. L'entropie favorise ainsi l'obtention de centroids flous. Dans ce travail nous avons comparé les méthodes CFE, FC* et FKM sur neuf jeux de données réelles.

2 Clustering avec centres flous et entropie

Soit $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ un ensemble de n objets catégoriques décrits par les attributs $A_1, A_2, \ldots, A_p$. Considérons pour chaque attribut catégorique A_l tel que $1 \leq l \leq p$ le domaine $DOM(A_l) = \{a_l^{(1)}, \ldots, a_l^{(n_l)}\}$ contenant des valeurs uniques. Ainsi $\mathbf{x}_i = [x_{i1}, \ldots, x_{il}, \ldots, x_{ip}]$ constitue un vecteur de p observations du $i^{ième}$ objet et x_{il} dénote la valeur du $l^{ième}$ attribut de l'objet $\mathbf{x}_i$. Soit k le nombre de classes et $\mathbf{v}_j$ les centres tel que $\mathbf{v}_j = [v_{j1}, \ldots, v_{jp}]$ pour tout $1 \leq j \leq k$. Les centres flous introduit dans (Kim et al., 2004) sont définis par l'utilisation d'un poids w pour chaque modalité associée à l'attribut A_l : $v_{jl} = [w_{jl}^{(1)} a_l^{(1)} \wedge \cdots \wedge w_{jl}^{(n_l)} a_l^{(n_l)}]$, avec $0 \leq w_{jl}^{(t)} \leq 1$ et $\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1$. La fonction objectif de CFE est définie par :

$$J_{CFE}(\mathbf{U}, \mathbf{W}) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m d(\mathbf{x}_i, \mathbf{v}_j) + \alpha \sum_{j=1}^{k} \sum_{l=1}^{p} \sum_{t=1}^{n_l} w_{jl}^{(t)} log(w_{jl}^{(t)})$$