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 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ un ensemble de n objets catégoriques décrits par les attributs $A_1, A_2, \dots, A_p$. Considérons pour chaque attribut catégorique A_l tel que $1 \le l \le p$ le domaine $DOM(A_l) = \{a_l^{(1)}, \dots, a_l^{(n_l)}\}$ contenant des valeurs uniques. Ainsi $\mathbf{x}_i = [x_{i1}, \dots, x_{il}, \dots, x_{ip}]$ 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 $\mathbf{v}_j = [v_{j1}, \dots, v_{jp}]$ pour tout $1 \le j \le k$. 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 : $v_{jl} = [w_{jl}^{(1)} a_l^{(1)} \wedge \dots \wedge w_{jl}^{(n_l)} a_l^{(n_l)}]$, avec $0 \le w_{jl}^{(t)} \le 1$ et $\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1$. La fonction objectif de CFE est définie par :

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

Avec $\mathbf{U} = (u_{ij})$ la matrice de partitionnement floue, $\mathbf{W}$ l'ensemble des poids, $m > 1$ le coefficient de paritionement flou, $d(\mathbf{x_i}, \mathbf{v}_j) = \sum_{l=1}^{p} \sum_{t, x_{il} \neq a_l^{(t)}}^{n_l} w_{jl}^{(t)}$ et $\alpha > 0$ un coefficient permettant de contrôler l'importance donnée à l'entropie. J_{CFE} est soumis aux contraintes $\sum_{j=1}^{k} u_{ij} = 1, \forall i\,;\, u_{ij} > 0\; \forall i, \forall j\,;\, 0 \leq w_{jl}^{(t)} \leq 1, \forall j, \forall l, \forall t$ and $\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1, \forall j, \forall l$.
Les formules de mises à jour des centres de FC* et CFE ainsi que leurs démonstrations sont décrites dans (Djiberou Mahamadou et al., 2020).

3 Résultats expérimentaux

Le tableau TAB. 1 présente les scores moyens (indice de rand) obtenus sur 100 itérations pour chaque méthode sur neuf jeux de données disponibles sur l'archive UCI. Les expériences sont conduites pour des valeurs de m variant de 1.1 à 2. Nous avons observé que m=1.2 donnait en général les scores optimaux pour tous les jeux de données considérées.
A partir du tableau ci-dessous, on remarque que les algorithmes CFE et FKM ont des scores

TAB. 1 – *Scores moyens obtenus pour m=1.2*

	FKM	FC*	CFE
Lung	0.55	0.56	0.56
Soybean	0.9	0.84	**0.93**
Zoo	0.88	0.87	0.9
Breast Cancer	0.52	0.5	0.**56**
Dermatology	0.48	0.42	**0.57**
Votes	**0.77**	0.76	**0.77**
Credits	**0.67**	0.6	**0.67**
Cars	**0.5**	0.48	**0.5**
Mushrooms	0.56	0.5	**0.58**

supérieurs à celui de FC*. On constate aussi que CFE et FKM ont pour certains jeux de données les mêmes scores. Des analyses statistiques montrent que sur tous les jeux de données considérées pour m variant de 1.1 à 2, CFE donne les meilleures performances. Tandis que la différence de performance est significative entre CFE et FC*, cette dernière est non significative pour CFE et FKM. Néanmoins, l'obtention des centres flous avec la méthode CFE à travers l'entropie peut permettre une meilleure caractérisation et interprétation des classes.

Références

Djiberou Mahamadou, A. J., V. Antoine, E. M. Nguifo, et S. Moreno (2020). Categorical fuzzy entropy c-means. In *Intl. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6.

Huang, Z. et M. Ng (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems 7*(4), 446–452.

Kim, D., K. Lee, et D. Lee (2004). Fuzzy clustering of categorical data using fuzzy centroids. *Pattern recognition letters 25*(11), 1263–1271.