CLUSTERING IMBALANCED FUNCTIONAL DATA

Michelle Carey [1], Catherine Higgins[1]

[1] School of Mathematics and Statistics, University College Dublin, (e-mail: michelle.carey@ucd.ie, catherine.higgins1@ucdconnect.ie)

ABSTRACT: Class imbalance is a common problem in functional clustering where some clusters have significantly more curves than other clusters. In such cases, most clustering algorithms tend to prioritize the majority class, resulting in sub-optimal cluster assignments. We propose a functional iterative hierarchical clustering approach to address the issue of class imbalance in functional data clustering. The performance of the proposed approach is compared with existing approaches. The proposed approach yields more accurate cluster assignments and a more precise approximation of the average trajectory of the curves within each cluster.

KEYWORDS: functional data, unsupervised clustering, class imbalance

1 Introduction

Unsupervised functional clustering techniques classify a sample of curves into homogeneous groups of curves, without prior knowledge of the true underlying clustering structure. The two common approaches for clustering functional data are: to obtain an approximation of the functional data in a finite-dimensional space and then use traditional clustering tools to cluster the resulting vectors (Chen *et al.*, 2012 and Wang & Xu, 2017) or to perform functional model-based clustering (Bouveyron & Jacques, 2011, Bouveyron *et al.*, 2015, and Centofanti *et al.*, 2023). See Jacques & Preda, 2014 for detailed reviews of functional clustering methods.

The problem of class imbalance occurs when the number of curves in one cluster significantly exceeds the number of curves in another cluster, posing a difficult challenge for most functional clustering algorithms. Minor clusters are often classified incorrectly into major clusters, which results in inaccurate cluster assignments and a poor approximation of the average trajectory of the curves within each cluster.

By extending Carey *et al.*, 2016 iterative hierarchical clustering method to a functional data context, we provide an approach for clustering imbalanced functional data.

2 Functional Iterative hierarchical clustering

The observations of the behavior of the curves at discrete points are subject to measurement error, that is $y_{i,j} = x_i(t_{i,j}) + \varepsilon_{i,j}$, where $t_{i,j}$ denotes the finite set of times from which one samples the i^{th} curve and the errors $\varepsilon_{i,j}$ are assumed to be independently distributed with mean 0 and a constant variance σ^2. Given the observed values $y_{i,j}$ for $i = 1,\ldots,N$ and $j = 1,\ldots,M_i$. The functional IHC algorithm performs the following steps:

1. **Reconstruct the functional form from the discrete observations**: Approximate the curves via a basis function expansion, that is, $\hat{x}_i(t) = \sum_{k=1}^{K} c_k \phi_k(t)$, and estimate the coefficients of the basis function expansion $\{c_k : k = 1,\ldots,K\}$ using the standard penalized least squares smoothing approach of Ramsay & Silverman, 2005.
2. **Cluster the first derivative of the curves:** The estimated first derivative of the curves evaluated at the points $\boldsymbol{t} = [t_{1,1},\ldots,t_{N,M}]$ are then given by the $N \times M$ matrix $\mathbf{A} = \sum_{k=1}^{K} \hat{c}_k D\phi_k(\boldsymbol{t})$, where $M = max(M_i)$ for $i = 1,\ldots,N$. Let α_{min} and α_{max} be the minimum and maximum of the Spearman rank correlation between all the possible pairs of the rows of $\mathbf{A}$. Define $[\alpha_{min},\ldots,\alpha_{max}]$, as a grid of Q equally spaced values from α_{min} to α_{max}. Cluster the rows of $\mathbf{A}$ using the iterative hierarchical clustering method proposed in Carey *et al.*, 2016. Select the optimal α_{opt} so that the value of the Davies-Bouldin index is minimized.

3 Simulations

The simulated sample curves X_i are realizations of a Gaussian process with the Matérn covariance function $C(s,t) = 0.2 \times \exp(-0.3\|s-t\|)$, over the domain $I = [0,15]$. To obtain six simulated groups of curves we define six different mean functions: $\sin(2\pi t)$, $\cos(2\pi t)$, $\sin(4\pi t + \pi/2)$, $\sin(4\pi t - \pi/2)$, $\sin(3\pi t + \pi/3)$ and $\sin(6\pi t - \pi)$. The six clusters are large, medium, and small in size, that is $N = 500, 500, 200, 15, 10, 3$. The sample data are given by, $Y_{i,j} = X_i(t_{i,j}) + \varepsilon_{i,j}$ for $i = 1,\ldots,N$ and $j = 1,\ldots,M_i$, where $\varepsilon_{i,j}$ is a normally distributed random variables with mean 0 and standard deviation σ_ε. We assume that $t_{i,j}$ are obtained from an equally spaced discretization of the domain and that this is the same for all curves.

The functional IHC is compared with the following eight state-of-the-art functional clustering methods: funFEM (Bouveyron *et al.*, 2015), funHDDC (Bouveyron & Jacques, 2011); SaS-Funclust (Centofanti *et al.*, 2023), functional EMCluster (Chen *et al.*, 2012), functional kCFC (Chiou & Li, 2007),

FADPclust1 and FADPclust2 (Wang & Xu, 2017). Table (1) presents the accuracy of the clustering methods measured by the average Adjusted Rand Index (μ_{ARI}) and the average Davies-Bouldin index (μ_{DB}). The adjusted rand index

Table 1. *The average Adjusted Rand Index (μ_{ARI}); and Davies-Bouldin index μ_{DB} for all eight functional clustering methods*

Method	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}
	M=15				M=200			
	$\sigma_\varepsilon = 0.05$		$\sigma_\varepsilon = 0.15$		$\sigma_\varepsilon = 0.05$		$\sigma_\varepsilon = 0.15$	
funIHC	1.00	0.84	0.99	0.86	0.99	0.82	0.99	0.89
funFEM	0.55	6.18	0.55	0.99	0.53	0.86	0.52	0.95
funHDDC	0.53	1.08	0.47	1.11	0.30	3.29	0.28	3.41
SaS-Funclust	0.53	1.08	0.53	0.97	0.35	3.67	0.15	3.11
Functional EMCluster	0.94	1.08	0.96	1.01	0.68	1.76	0.69	1.90
Functional kCFC	0.97	1.08	0.90	1.52	0.54	1.64	0.64	2.10
FADPclust1	0.68	1.66	0.71	1.14	0.69	1.10	0.71	1.15
FADPclust2	0.68	1.02	0.68	1.16	0.74	0.94	0.71	1.06

ranges from 0 to 1 and measures the similarity between the clustering assignment and the true group structure. Clustering assignments are more accurate when the value is larger. The Davies-Bouldin index is based on the ratio of within-cluster distances to between-cluster distances. Clusters that are farther apart and less dispersed will result in a lower index. The funIHC obtains the highest adjusted rand index and the lowest Davies-Bouldin index. FunIHC is the only approach to correctly identify the number of curves in each cluster and the true average temporal pattern. FunFEM, FunHDDC, Sasfunclust, Functional EMCluster, and Functional kCFC provide a good approximation of the average temporal patterns for the larger clusters but provide a poor approximation for ($N < 200$). FADPclust1 and FADPclust2 miss-classifies the small and medium clusters into the larger clusters resulting in poor approximations of the average temporal pattern for all clusters.

4 Conclusion

A functional iterative hierarchical clustering approach is proposed that can effectively address the issue of class imbalance in functional data clustering. The proposed approach is shown to outperform existing approaches in terms of the

accuracy in the cluster assignments and the approximations of the average temporal pattern of the cluster members.

References

BOUVEYRON, CHARLES, & JACQUES, JULIEN. 2011. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5**(4), 281–300.

BOUVEYRON, CHARLES, CÔME, ETIENNE, & JACQUES, JULIEN. 2015. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, **9**(4), 1726–1760.

CAREY, MICHELLE, WU, SHUANG, GAN, GUOJUN, & WU, HULIN. 2016. Correlation-based iterative clustering methods for time course data: the identification of temporal gene response modules for influenza infection in humans. *Infectious Disease Modelling*, **1**(1), 28–39.

CENTOFANTI, FABIO, LEPORE, ANTONIO, & PALUMBO, BIAGIO. 2023. Sparse and smooth functional data clustering. *Statistical Papers*, 1–31.

CHEN, WC, MAITRA, R, & MELNYKOV, V. 2012. EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution. *R Package, URL http://cran. r-project. org/package= EMCluster*.

CHIOU, JENG-MIN, & LI, PAI-LING. 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(4), 679–699.

DAVIES, DAVID L, & BOULDIN, DONALD W. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 224–227.

JACQUES, JULIEN, & PREDA, CRISTIAN. 2013. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, **112**, 164–171.

JACQUES, JULIEN, & PREDA, CRISTIAN. 2014. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 231–255.

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional Data Analysis*. Springer.

RODRIGUEZ, ALEX, & LAIO, ALESSANDRO. 2014. Clustering by fast search and find of density peaks. *science*, **344**(6191), 1492–1496.

WANG, XIAO-FENG, & XU, YIFAN. 2017. Fast clustering using adaptive density peak detection. *Statistical methods in medical research*, **26**(6), 2800–2811.