Clustering of Functional Data via Ensemble of Random Projections

Clustering of functional data via ensemble of Random Projections

Matteo Mori[1] and Laura Anderlucci[2]

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy
matteo.mori8@unibo.it
[2] Department of Statistical Sciences, University of Bologna, Bologna, Italy

Abstract. Clustering functional data presents challenges due to its infinite-dimensional nature. A common strategy to represent functions in a finite space is to consider basis function decomposition, where each curve is represented by a linear combination of basis functions; however, the need for accurate representation may still lead to high-dimensional embedding, making traditional clustering methods inadequate. In this work, we propose using Random Projections (RP) to reduce dimensionality while approximately preserving distances. Specifically, we propose to employ an ensemble approach that aggregates multiple clustering solutions from different projections, enhancing stability and accuracy. Its capability of recovering the true cluster membership is illustrated on a real data application.

Keywords: clustering, functional data, ensemble, random projections

1 Introduction

Given a set of heterogeneous observations with no prior information on group labels, cluster analysis aims at grouping similar observations into homogeneous groups. Many different methods have been proposed in the literature, according to different ways of defining and interpreting the concept of "similarity" between observations. For instance, model-based clustering methods assume that data come from a mixture of different (often Gaussian) distributions [2].

Functional data are a type of data that is gaining increasing importance in applications nowadays. It consists of infinite dimensional observations or curves that can be thought as realizations of a stochastic process $\{X(t) : \ t \in S\}$ over an interval S. A nice review of clustering methods for functional data is given in [12].

The infinite dimensionality nature of functional data makes clustering a difficult task. In particular, model-based clustering methods are not directly applicable: before employing traditional methods, the problem needs to be rephrased into a finite dimensional one; see e.g. [4].

There are many ways to represent functions in a finite space. In this paper we focus on basis function decomposition, where each curve is represented by a linear combination of basis functions. Usual choices for basis functions include Fourier

(if observations have a periodic behavior) or B-spline ones, among others. An exhaustive theoretical discussion on functional data representation can be found in [14]. Using basis functions rather than observed points allows to regularize the curves, so to account for measurement errors. However, in order to have an accurate representation of the data, basis function expansions can still lead to a high dimensional parameter space, which prevents from using traditional methods such as mixture models, due to the well-known curse of dimensionality.

Numerical problems can be overcome by resorting to regularized or to parsimonious models, at the expense of model flexibility. Alternatively, one can resort to some dimension reduction methods. Main approaches in this area are feature selection, where it is assumed that only a subset of the variables is relevant in uncovering the clusters structure, or feature extraction, where the clustering space is found by considering new artificial features, e.g. linear combinations of the original ones, like for instance Principal Components. Our idea is to use Random Projections (RP) to reduce the dimensionality of the data. This fairly recent approach consists in mapping at random the original high-dimensional data onto a lower subspace by using a random projection matrix. The Johnson-Lindenstrauss (JL) Lemma guarantees that pairwise distances between units are nearly preserved [13]. More formally, the lemma states that, given $x_1, \ldots, x_n \in \mathbb{R}^p$, $\epsilon \in (0, 1)$ and $d > 8\log(n)/\epsilon^2$, there is a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ such that

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

for all $i, j = 1, \ldots, n$.

However, the main drawback of employing RPs for cluster analysis is their variability: different projections may or may not highlight a grouping structure. In order to overcome this issue, the results on several different projections can be combined in an ensemble (for an application in the supervised framework see [5]).

Random projection ensembles have been successfully applied in the clustering of high dimensional data. For example, [1] proposed to apply model-based clustering onto a set of randomly projected data; the final clustering is obtained by aggregating the results of selected projections.

Inspired by work in [1], in this paper we propose to address the clustering of functional data by resorting to an ensemble of clustering solutions obtained from applying model-based clustering procedures on a lower dimensional embedding of the data. The proposal performance has been evaluated in an extensive simulation study; here, results on a real data application are presented.

2 Outline of the proposed method

When dealing with functional data, the observations are supposed to belong to an infinite dimensional space, whereas in practice one only has sampled curves observed into a finite set of observed points. The most common solution to this problem is to consider that sample paths belong to a finite dimensional space

spanned by some basis of functions (see, for example, [14]). Specifically, each curve $x_i,\ i \in \{1,\dots,N\}$ is considered as observed on the same time points $1,\dots,n$ and in the same interval. If the curve is observed without error, the linear combination

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) = \mathbf{c}_i'\boldsymbol{\phi}(t)\ ,$$

where $\mathbf{c}_i$ is the K-vector of coefficients for curve i and $\boldsymbol{\phi}$ is the functional vector whose elements are the basis functions ϕ_k (identical for all $i \in \{1\dots N\}$), has least squares solution

$$\hat{\mathbf{c}}_i = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{y}_i\ ,$$

where matrix $\boldsymbol{\Phi}$ in $\mathbb{R}^{n\times K}$ contains the values of the basis functions at the n sampling points and $\mathbf{y}_i$ is the vector of the observed values.

In order to account for measurement errors, the sum of squared errors can be penalized via a roughness penalty. A common choice for the penalty is

$$\int [D^m x(s)]^2 ds = \mathbf{c}'\left[\int D^m\boldsymbol{\phi}(s) D^m\boldsymbol{\phi}'(s) ds\right]\mathbf{c}\ ,$$

which is often taken to be with respect to the second derivative (i.e. $m = 2$). In practice, more general roughness penalties could be used.

Thus, the new function to minimize is

$$(\mathbf{y}_i - \boldsymbol{\Phi}\mathbf{c}_i)'(\mathbf{y}_i - \boldsymbol{\Phi}\mathbf{c}_i) + \lambda\mathbf{c}_i' R\mathbf{c}_i\ ,$$

where $R = \int D^m\boldsymbol{\phi} D^m\boldsymbol{\phi}'$, from which the expression for the estimated coefficient vector is

$$\hat{\mathbf{c}}_i = (\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda R)^{-1}\boldsymbol{\Phi}'\mathbf{y}_i\ .$$

The number of basis functions K and the roughness penalty λ can be set or chosen over a grid of values via Generalized Cross-Validation (GCV).

Once the matrix $C \in \mathbb{R}^{N\times K}$, containing the coefficients of the basis functions for each curve, is recovered, we propose to generate B random projections matrices $A_1,\dots,A_B$, with $A_b \in R^{K\times d}$, $d \ll K$, $b = 1,\dots,B$. The high dimensional matrix C can then be embedded in the lower dimensional spaces $X_b = CA_b$. A set of clustering solutions are obtained by applying unsupervised classification methods, here mixtures of multivariate normal densities, to the dimensionally reduced data X_b, $b = 1,\dots,B$. There are several kinds of projection matrices A_b that satisfy the JL's Lemma; most common examples include the Haar projection matrix and the Gaussian one, among others. In our experimental studies, we have considered both and results look pretty similar; due to space limitation, only those of Gaussian's projection matrices are reported.

In order to discard the non-informative projections, the clustering results can be sorted according to a cluster quality measure. The best B^* solutions are then aggregated via consensus clustering [7], so as to obtain a single cluster membership vector.

The dimension of the projected space d, the number of projections B, the size of the ensemble B^* and the cluster quality measures are parameters that need to be set.

3 Real data application: Olive Oil Data

We assess the proposed method on a real dataset containing samples of olive oil analyzed by spectroscopy [8]. The data consist of $n = 138$ Greek olive oil samples, equally divided in three classes: 46 observations are authentic extra virgin olive oils, 46 samples are adulterated by the addition of 1% sunflower oil, and the last 46 were contaminated by adding 5% sunflower oil. The spectroscopy of each unit is recorded over $p = 1050$ wavelengths.

The goal is to uncover the grouping structure of units, with particular interest in identifying the unadulterated oils and the contaminated ones.

The optimal combination of parameters returned by GCV is $K = 1050$ basis and $\lambda = 10^{-5}$ (the considered penalty is on the second derivative, i.e. $m = 2$).

Data according to the obtained basis expansions is depicted in Figure 1, where different colors identify different classes.

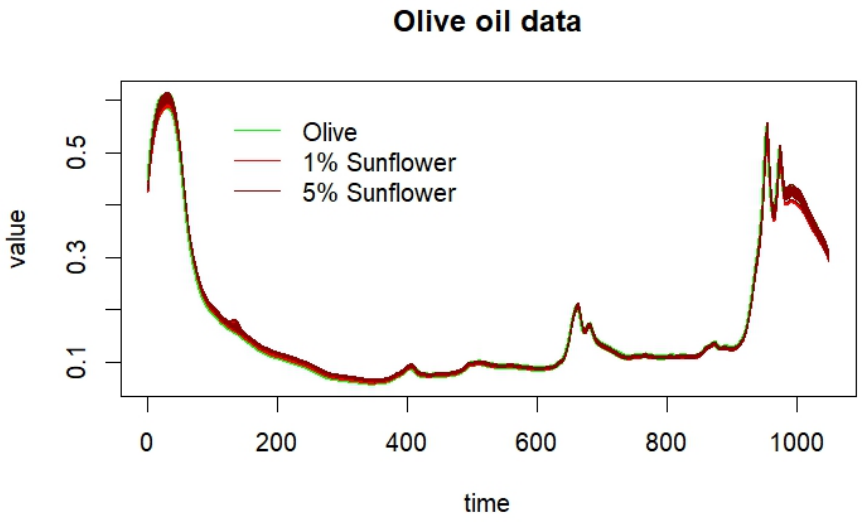

Fig. 1. Oil data spectra (functional basis representation)

Dasgupta in [6] proved that data from an arbitrary mixture of Gaussian distributions can be randomly embedded into a subspace of just $\mathcal{O}(\log G)$ dimensions, while preserving the group structure almost perfectly. Furthermore, when $d < \log G$, the worsening of the mapping performance is gradual. Empirical studies described in [1] suggest to set the dimension of the projected subspace d to $\lfloor 10log(G) \rceil + 1$, where G is the number of groups, considered as known; they observed how higher values of d do dot noticeably improve the final performance.

When G is not known, the choice of d can be based on the expected or desired number of groups. Following [1], we set $\lfloor 10log(G) \rceil + 1 = 12$. Overall $B = 1000$ projections were generated and the best 28 ones were selected and combined in the ensemble [10].

Since the cluster labels were available, the adjusted Rand Index (ARI) was employed as a measure of clustering recovery capability. In order to evaluate the performance of the proposal, we compare our procedure (RPClust_FD) with the several reference methods in the literature, some accounting for the functional structure of the data. Specifically:

1. K-means [9] on the original dataset (KM orig).
2. Gaussian Mixture Models [15] on the original dataset (GMM orig).
3. High-dimensional data clustering (HDDC) [3] on the original data, with subspace dimension selected via BIC (HDDC_B).
4. HDDC on the original data, with subspace dimension selected via Cattell's scree-test (HDDC_C).
5. K-means on the basis coefficients (KM fun).
6. Gaussian Mixture Models on basis coefficients (GMM fun).
7. funHDDC [4], a functional version of HDDC (funHDDC).
8. Mixture model based on functional random variables density approximation [11] (funclust).

For the methods that consider the functional nature of the data (namely 5, 6, 7 and 8), the choice of an appropriate basis representation is crucial. For a fair comparison, we ran each method with the values of K and λ obtained by GCV. Table 1 reports the ARI of the considered methods and the corresponding computational time. The proposal (RPClust_FD), together with HDDC_B (that however does not account for the functional nature of the data), obtained the highest ARI; the confusion matrix reported in Table 2 shows that the method is able to distinguish clearly between the authentic oil from the adulterated ones.

Table 1. Adjusted Rand Index (ARI) and computational time (in seconds) for the olive oil data.

Method	ARI	time (s)
1. KM orig	0.24	0.46
2. GMM orig	0.30	0.64
3. HDDC_B	0.70	1.77
4. HDDC_C	0.10	0.91
5. KM fun	0.24	6.94
6. GMM fun	0.30	7.25
7. funHDDC	0.54	7469.96
8. funclust	0.00	41.40
9. RPClust_FD	0.70	76.10

Table 2. Confusion matrix of the clustering results of RPClust_FD and the true labels.

		Oil label		
		0%	1%	5%
Cluster	1	0	16	45
	2	0	30	1
	3	46	0	0

4 Conclusion

The real data application shows that resorting to ensemble of Random Projections can be a valid tool to deal with high dimensionality while providing a data oblivious method for obtaining informative subspaces in the context of functional data. The computational time is comparatively longer that alternative approaches but it can be efficiently parallelized. In this work, the number of clusters is considered as fixed and known; methods to optimally choose G is object of future developments.

References

1. Anderlucci, L., Fortunato, F., Montanari, A.: High-Dimensional Clustering via Random Projections. J. Classif. 39, 191–216 (2022). doi:10.1007/s00357-021-09403-7.
2. Banfield, J. D., Raftery, A. E.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics. 49(3), 803–821 (1993). doi:10.2307/2532201.
3. Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Comput. Stat. Data Anal. 52(1), 502-519 (2007). doi:10.1016/j.csda.2007.02.009.
4. Bouveyron, C., Jacques, J.: Model-based Clustering of Time Series in Group-specific Functional Subspaces. Adv. Data Anal. Classif. 5, 281-300 (2011). doi:10.1007/s11634-011-0095-6.
5. Cannings, T. I., Samworth, R. J.: Random-Projection Ensemble Classification. J. R. Stat. Soc. B. 79(4), 959–1035 (2017). doi:10.1111/rssb.12228.
6. Dasgupta, S.: Experiments with random projections. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI '00), 143–151 (2000). doi:10.5555/2073946.2073964.
7. Dimitriadou, E., Weingessel, A., Hornik, K.: A Combination Scheme for Fuzzy Clustering. In: Pal, N.R., Sugeno, M. (Eds.), Advances in Soft Computing — AFSS 2002, vol. 2275. Springer, Berlin, Heidelberg (2002). doi:10.1007/3-540-45631-7_44.
8. Downey, G., McIntyre, P., Davies, A. N.: Detecting and Quantifying Sunflower Oil Adulteration in Extra Virgin Olive Oils from the Eastern Mediterranean by Visible and Near-Infrared Spectroscopy. J. Agric. Food. Chem. 50(20), 5520-5525 (2002). doi:10.1021/jf0257188.
9. Hartigan, J. A., Wong, M. A.: Algorithm AS 136: A K-means clustering algorithm. J. R. Stat. Soc. C. 28, 100–108 (1979). doi:10.2307/2346830.
10. Hornik, K.: A CLUE for CLUster Ensembles. J. Stat. Softw. 14(12), 1–25 (2005). doi:10.18637/jss.v014.i12.
11. Jacques, J., Preda, C.: Funclust: A curves clustering method using functional random variables density approximation. Neurocomputing. 112, 164–171 (2013). doi:10.1016/j.neucom.2012.11.042.
12. Jacques, J., Preda, C.: Functional Data Clustering: A Survey. Adv. Data Anal. Classif. 8, 231-255 (2013). doi:10.1007/s11634-013-0158-y.
13. Johnson, W. B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. 26, 1, 189-206 (1984). doi:10.1090/conm/026/737400.
14. Ramsay, J. O., Silverman, B. W.: Functional Data Analysis. Springer (2005).
15. Scrucca, L., Fraley, C., Murphy, T. B., Raftery, A. E.: Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman & Hall/CRC (2023). https://mclust-org.github.io/book/.