Clustering Data Streams via Functional Data Analysis: a Comparison between Hierarchical Clustering and K-means Approaches

(*Classificazione di Data Stream con l'Analisi dei Dati Funzionali: un Confronto tra Cluster Gerarchica e Metodo delle K-medie*)

Fabrizio Maturo, Francesca Fortuna, and Tonio Di Battista

Abstract Recently, the analysis of web data, has become essential in many research fields. For example, for a large number of companies, corporate strategies should be based on the analysis of customer behaviour in surfing the world wide web. The main issues in analysing web traffic and web data are that they often flow continuously from a source and are potentially unbounded in size, and these circumstances inhibit to store the whole dataset. In this paper, we propose an alternative clustering functional data stream method to implement existing techniques, and we address phenomena in which web data are expressed by a curve or a function. In particular, we deal with a specific type of web data, i.e. trends of google queries. Specifically, focusing on top football players data, we compare the functional k-means approach to the functional Hierarchical Clustering for detecting specific pattern of search trends over time.

Abstract Recentemente, l'analisi dei dati web é diventata essenziale in molti campi di ricerca. Informazioni sulle ricerche nei motori di ricerca, sul numero di visite, sul tempo trascorso sul sito web, sulla provenienza degli utenti e sul successo di una pubblicitá online, sono essenziali per prevedere vendite future, analizzare le performance passate e monitorare i modelli di acquisto. Tuttavia, lo studio del traffico web presenta alcuni problemi metodologici in quanto questi dati fluiscono continuamente inibendone la totale archiviazione ed ostacolando l'applicazione di tecniche di analisi standard. Questo articolo propone un metodo alternativo di classificazione rispetto a quelli noti in letteratura affrontando situazioni in cui i dati web sono

Fabrizio Maturo
"G. d' Annunzio" University, DEA, Pescara, Italy, e-mail: f.maturo@unich.it

Francesca Fortuna
"G. d' Annunzio" University, DISFPEQ, Pescara, Italy, e-mail: francesca.fortuna@unich.it

Tonio Di Battista
"G. d' Annunzio" University, DISFPEQ, Pescara, Italy, e-mail: dibattis@unich.it

espressi da una curva o da una funzione. L'analisi é condotta su un tipo specifico di dati web, cioé l'andamento delle ricerche su Google. Nello specifico, concentrandoci sui dati dei migliori giocatori di calcio, l'obiettivo é quello di confrontare l'approccio delle k-medie con quello della cluster gerarchica per dati funzionali al fine di rilevare modelli di comportamento nelle tendenze di ricerca attraverso il tempo.

Key words: clustering football players, data streaming, google query, google trends, FDA

1 Introduction

In recent decades, the analysis of web-data has attracted the attention of many companies dealing with marketing and advertising in the Sports sector, and particularly in the football industry. Due to the massive diffusion of high speed network and internet technology, the analysis of consumers' behavior in surfing the web has become a fundamental data for understanding consumer preferences. Data collected on the web are referred to as data streams, whose main characteristics are that they continuously flow. Since data streams are potentially infinite, the identification of common patterns through clustering techniques become a crucial issue. However, clustering methods suffer from many problems related to the nature of the data, such as clusters robustness over time, the difficulty of exploring data at different portions of the stream, and the computational time-consuming issue. Because each stream can be seen as a function in a continuous domain, i.e. time, this paper proposes to analyze web-data through the functional data analysis (FDA) approach [Ramsay and Silverman, 2005] and discusses the benefits of this method focusing on the clustering data streams problem. The main advantage provided by FDA approach is the drastic dimensionality reduction of data streams, which is gained by converting discrete observations into functions. Moreover, the use of additional functional tools may sometimes reveal crucial information better than the original data [Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Maturo and Di Battista, 2018, Fortuna and Maturo, 2018, Maturo, 2018, Maturo et al., 2018]. Specifically, in this paper, the FDA approach is proposed for clustering data streams using either hierarchical clustering and K-means approaches. In this study, we focus our attention on a semimetric based on the functional principal components (FPCs) decomposition for both clustering technique. The methodological approach is implemented for analyzing a real data set concerning the Google queries regarding 20 top football players. The final aim of this contribution is to propose the use of functional clustering methods for the analysis of web-queries, and to compare the results of the main two functional clustering approaches.

2 Materials and Methods

The basic idea behind a functional interpretation of web-data is that each stream is a sample from a smooth function of time. Because in real applications curves are observed at a finite set of sampling points, the functional form of a stream, say $X(t)$, can be reconstructed and represented by a basis expansion as follows: $X_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t) \quad i = 1, ..., n$ where $X_i(t)$ is the reconstructed function for the i-th unit; $\phi_k(t)$ are linearly independent and known basis functions; and a_{ik} are coefficients that link each basis function together in the representation of $X_i(t)$. Various basis systems can be adopted, depending on the characteristics of curves [Ramsay, 1991]. In this work, we consider the least squares approximation with B-splines basis for the functional representation of web-data. To achieve an optimal representation of curves into a function space of reduced dimension, the functional principal component analysis (FPCA) can be adopted [Ferraty and Vieu, 2006]. In particular, let us assume that the observed curves are centered so that the sample mean is equal to zero. Then, for each unit ($i = 1, .., n$), the j-th principal component score is given by $\xi_{ij} = \int_T x_i(t) f_j(t)\, dt$, where the weight functions or loadings $f_j(t)$ are the solutions of the eigenequation where $c(t,s)$ is the sample covariance function and $\lambda_j = Var[\xi_j]$ [Aguilera and Aguilera-Morillo, 2013]. Then, the sample curves admit the following principal component decomposition: $x_i(t) = \sum_{j=1}^{p} \xi_{ij} f_j(t) \quad i = 1, ..., n$ where p denotes the total number of functional principal components. By truncating this representation in terms of the first q principal components ($q << p$), we can obtain an approximation of the sample curves whose explained variance is given by $\sum_{j=1}^{q} \lambda_j$. If we assume that the observed functions are expressed in terms of B-splines, then the weight functions $f_j(t)$, admit the following basis expansion: $f_j(t) = \sum_{k=1}^{K} b_{jk}\phi_k(t), \; j = 1, 2, ..., q$. To identify specific common patterns of data-streams, clustering of functions is carried out in combination with dimension reduction in order to remove the effect of irrelevant functional information. In particular, the distance among functional data is computed using the FPCA method according to the following semi-metric [Ferraty and Vieu, 2006]: $d^{(q)}\left(X_i(t), X_{i'}(t)\right) = \left[\sum_{j=1}^{q} \left(\xi_{ij} - \xi_{i'j}\right)^2 \left\| f_j^{(q)} \right\|\right]^{1/2} \quad i \neq i'$ where $\left\| f_j^{(q)} \right\| = \int_T f_j(t)^2\, dt$, and q denotes the reduced dimensional space at q components [Ferraty and Vieu, 2006, Febrero-Bande and de la Fuente, 2012], chosen according to the criterion of their explained variability.

The basic idea of this unsupervised clustering approach is to find a partition for which the variability within clusters is minimized. The most used algorithm, in this context, is the k-means. Starting from n functional observations, this method aims to assembly units into $G \leq n$ groups, $C_1, C_2, ..., C_G$ so as to minimize the within-cluster sum of squares. The first step of this iterative procedure consists in fixing G initial centroids, $\psi_1^{(0)}(t), ..., \psi_G^{(0)}(t)$. Then, each function is assigned to the cluster whose centroid, at the previous iteration $(m-1)$ is the nearest according to the chosen distance. Once all the functions have been assigned to a cluster, the cluster means are

updated as follows: $\psi_g^m(t) = \sum_{x_i(t) \in C_g} \frac{x_i(t)}{n_g}$ where n_g is the number of functions in the g-th cluster, C_g.

Despite this approach could be extended to first and second derivatives or other functions derived from the original ones [Fortuna et al., 2018, Fortuna and Maturo, 2018], in this context, we limit our attention to the original b-spline approximation. Effectively, the aim of this paper is to propose thise approach and compare its results to the agglomerative hierarchical method. In this setting, the classification strategy consists of a series of partitions, which may run from a single cluster containing all the functions (divisive methods), to n clusters, each containing a single function (agglomerative methods). In order to determine which groups should be merged (for agglomerative approach) or divided (for divisive approach), different metrics and linkage methods can be used. In this context, the agglomerative approach with the average linkage method is used. In addition, the gap statistic [Tibshirani et al., 2001] and silhouette plot are adopted for estimating the optimal number of clusters [Kassambara and Mundt, 2017].

3 Application

The proposed method has been applied to the number of queries collected by Google trends over two months (from the beginning of January 2018 to the end of February 2018) regarding the 19 football players with highest market values ($\geq$50 million euros), valued on December 19, 2017 (https://www.transfermarkt.it/); and participation in 2017-2018 UEFA Champions League. Table 1 shows the full list of statistical units considered in the study. Figures 1 and 2 illustrate top player queries over time and the corresponding spline approximation, respectively. Figure 3 presents the functional principal components decomposition: the first three FPCs explain 56.93%, 16.11%, and 8.99% of the total variability, respectively. Figure 4 exhibits the results of the k-means functional clustering of the 19 top players' queries over time whereas Figures 5 and 6 presents the hierarchical clustering findings. According to the gap statistic 7 and silhouette analysis 8, the optimal number of groups is two. Effectively, Figure 4 strongly highlights the presence of two distinct groups: the groups n.1 is characterized by higher variability and elevated levels of queries over time; conversely, the group n.2 presents low variability over time and a minor average level of queries. Table 1 specifies in detail the group membership according to the two different methods. 16 soccer players are distinguished by the same pattern of groups' membership. Only Ronaldo, Hazard, and Kroos change group in passage from k-means to hierarchical clustering.

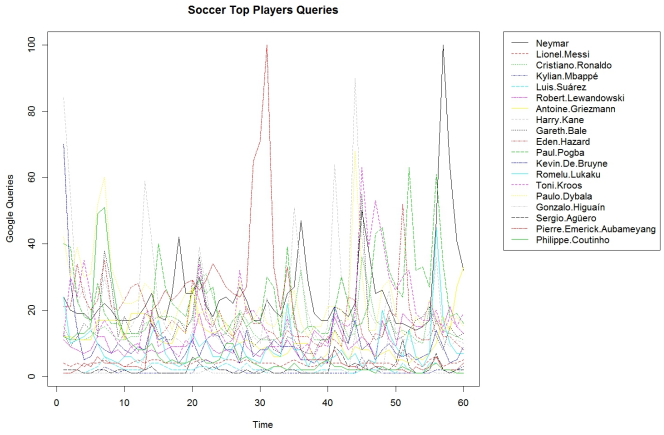

Fig. 1: Top Player Queries Over Time

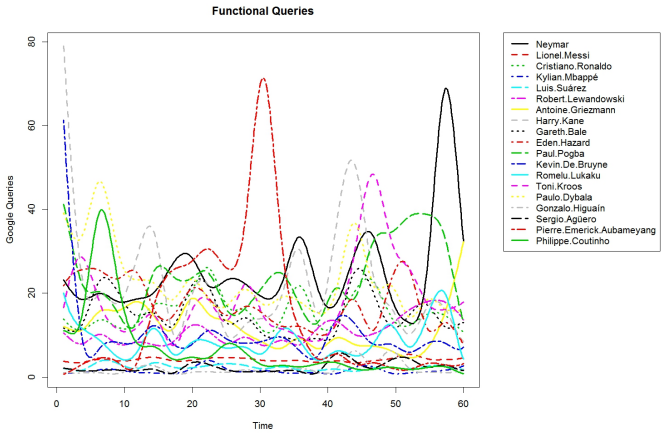

Fig. 2: Spline Approximation of Top Player Queries Over Time

4 Conclusion

The basic idea of this study is to use FDA to analyse datastreaming and in particular the number of queries on search engines. Hence, the originality of this research lies in the application, and in particular we have focused on cluster analysis using the two best known methods: the k-means and hierarchical clustering. Naturally, starting from this idea, many possible developments can be done in the field of functional regression, prediction, and supervised classification. In future research, it would be interesting to analyze in depth what are the reasons that lead to conflicting results between the two types of clustering methods. To achieve this goal, we should

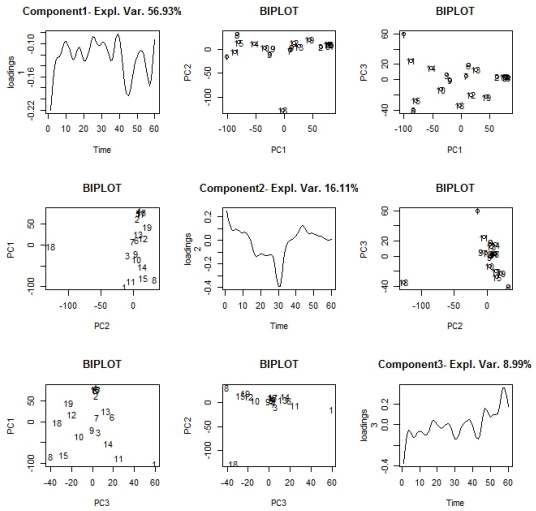

Fig. 3: Functional Principal Components Decomposition of Spline Approximation of Top Player Queries Over Time

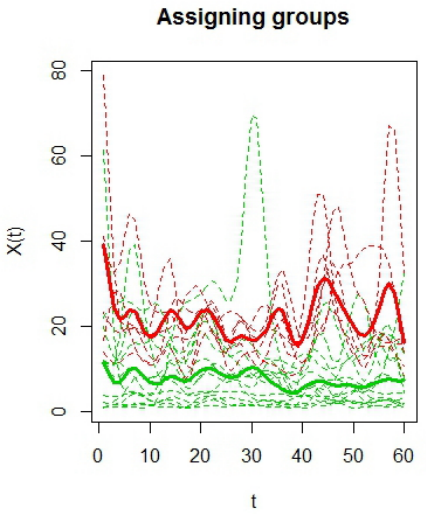

Fig. 4: K-Means Clustering of Top Player Queries Over Time (Red=Group n.1, Green=Group n.2)

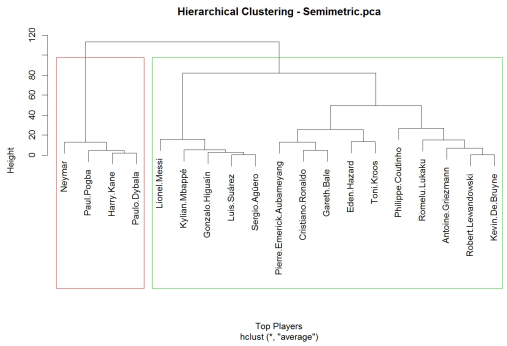

Fig. 5: Hierarchical Clustering of Top Player Queries Over Time (Red=Group n.1, Green=Group n.2)

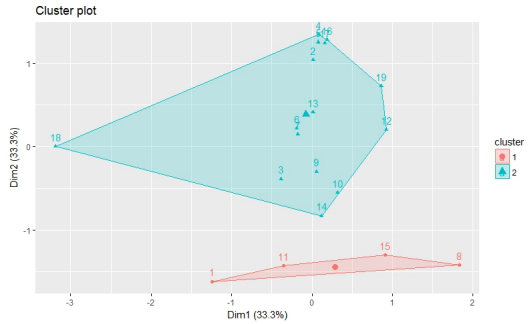

Fig. 6: Plot of Hierarchical Clustering of Top Player Queries Over Time

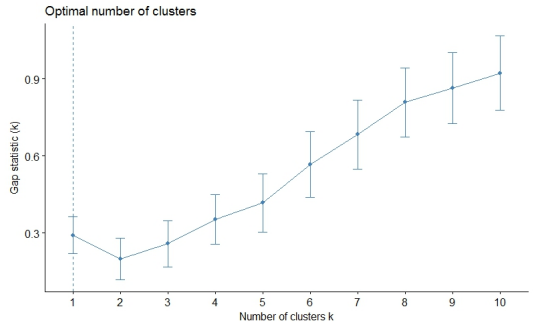

Fig. 7: Gap Statistic for Selecting Groups Number

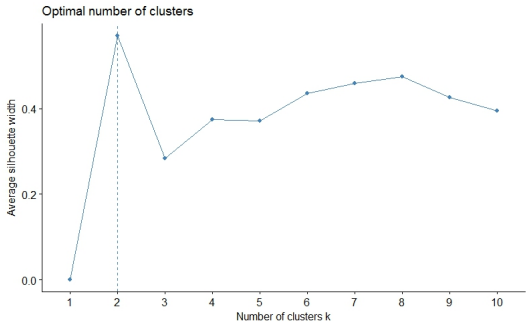

Fig. 8: Silhouette Analysis for Selecting Groups Number

No.	Player name	Age	Football club	Group (K-Means)	Group (Hierarchical)
1	Neymar	25	FC Paris Saint-G.	1	1
2	Messi	30	FC Barcelona	2	2
3	Ronaldo	30	Real Madrid CF	1	2
4	Mbappé	18	FC Paris Saint-G.	2	2
5	Suárez	30	FC Barcelona	2	2
6	Lewandowski	29	FC Bayern Monaco	2	2
7	Griezmann	26	Atlético de Madrid	2	2
8	Kane	24	Tottenham Hotspur	1	1
9	Bale	28	Real Madrid CF	2	2
10	Hazard	26	Chelsea FC	1	2
11	Pogba	24	Manchester United	1	1
12	De Bruyne	26	Manchester City	2	2
13	Lukaku	24	Manchester United	2	2
14	Kroos	27	Real Madrid CF	1	2
15	Dybala	24	Juventus FC	1	1
16	Higuaín	30	Juventus FC	2	2
17	Agüero	29	Manchester City	2	2
18	Aubameyang	28	Borussia Dortmund	2	2
19	Coutinho	25	FC Liverpool	2	2

Table 1: Group Membership of 19 Football Top Players according to two functional clustering methods.

also draw this study on the use of further semi-metrics, e.g. based on first and second derivatives, and different types of hierarchical clustering approaches, e.g. single-linkage and complete-linkage.

References

A. Aguilera and M. Aguilera-Morillo. Penalized PCA approaches for b-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219 (14):7805–7819, mar 2013. doi: 10.1016/j.amc.2013.02.009.

M. Febrero-Bande and M. de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software, Articles*, 51(4): 1–28, 2012. doi: 10.18637/jss.v051.i04.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer, New York, 2006.

F. Fortuna and F. Maturo. K-means clustering of item characteristic curves and item information curves via functional principal component analysis. *Quality & Quantity*, mar 2018. doi: 10.1007/s11135-018-0724-7.

F. Fortuna, F. Maturo, and T. Di Battista. Unsupervised classification of soccer top players based on google trends. *Quality and Reliability Engineering International*, 2018. doi: 10.1002/QRE-17-0561.

A. Kassambara and F. Mundt. factoextra: Extract and visualize the results of multivariate data analyses. 2017. URL `https://cran.r-project.org/web/packages/factoextra/index.html`.

F. Maturo. Unsupervised classification of ecological communities ranked according to their biodiversity patterns via a functional principal component decomposition of hill's numbers integral functions. *Ecological Indicators*, 90:305–315, jul 2018. doi: 10.1016/j.ecolind.2018.03.013.

F. Maturo and T. Di Battista. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. *Ecological Indicators*, 84(C):70 – 81, 2018. doi: 10.1016/j.ecolind.2017.08.016.

F. Maturo, F. Fortuna, and T. Di Battista. Testing equality of functions across multiple experimental conditions for different ability levels in the IRT context: The case of the IPRASE TLT 2016 survey. *Social Indicators Research*, apr 2018. doi: 10.1007/s11205-018-1893-4.

J. Ramsay. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630, 1991.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, 2nd edn*. Springer, New York, 2005.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, may 2001. doi: 10.1111/1467-9868.00293.