Recent advances in functional data stream classification

Tonio Di Battista
DiSFPEQ, University G. D Annunzio, Chieti-Pescara, Italy, - dibattis@unich.it

Francesca Fortuna
DiSFPEQ, University G. D Annunzio, Chieti-Pescara, Italy - francesca.fortuna@unich.it

Fabrizio Maturo
Department of Business Administration, University G. D Annunzio, Chieti-Pescara, Italy - f.maturo@unich.it

Abstract

High frequency processing and streaming data analysis have recently attracted the attention of statistical researchers. The stream processing model assumes that data flow continuously from a source and that they are potentially unbounded in size. This circumstance inhibits to store the whole dataset. Moreover this kind of data is useful when the analytics need to be done in real time. In fact, in some fields the value of the analysis decreases with time. The analyst cannot reanalyze the data after it is streamed, so it is important to use appropriate tools. In this paper we propose an alternative functional data stream classification to implement existing techniques and we address phenomena in which the statistical units are expressed through a process defined by a curve or a function. In particular, we propose a general theoretical methodology when units belong to a non convex functional space.

Keywords: clustering; functional data analysis; convex spaces.

1. Introduction

Data mining is a process of extracting useful information from large volume of database; it is the practice of automatically searching large store of data to discover patterns and trends that go beyond simple analysis. Data mining process is called "discovery" of looking in a data warehouse to find hidden patterns without a predetermined idea about what patterns may be. So, data mining is also known as a knowledge discovery in data (KDD). Data mining process has two major functions: classification and clustering. Data stream mining is a process of extracting knowledge structure from continuous, rapid data records. It can be considered as a subfield of data mining. The characteristics of data streaming are that data continuously flow, data size is extremely large and potentially infinite. Thus, data stream is an appropriate model when a large volume of data is arriving continuously and it is either unnecessary or impractical to store the data in some form of memory. Some applications naturally generate data streams as opposed to simple data sets. In the data stream model, the data points can only be accessed in the order in which they arrive. Random access to the data is not allowed; memory is assumed to be small relative to the number of points, and so only a limited amount of information can be stored. Example of data stream includes: web searches, computer network traffic, phone conversation, ATM transaction and sensor data. The challenge facing algorithm designers is to perform meaningful computation with these restrictions. A common form of data analysis in these applications involves clustering, i.e., partitioning the data set into subsets (clusters) such that members of the same cluster are similar and members of distinct clusters are dissimilar (O'Callaghan et al., 2002). Data stream classification is a challenging problem because of two important properties of the stream: its infinite length and evolving nature namely, concept drift and concept evolution. Concept drift occurs when the class labels of a set of examples change over time. Concept evolution occurs when a new class emerges in the stream. However, both concept drift and concept evolution may also occur simultaneously. In any case, the challenge is to build a classification model that is consistent with the current concept (Masud et al., 2011). One of the main fields of application of data streaming is the web traffic. Web sites receive streams of various types. For example, Google receives several hundred million search queries per day. Yahoo! accepts billions of "clicks" per day on its various sites. Many interesting things can be learned from these streams: first a sudden increase in the click rate for a link could indicate a virus attack or it could mean that the link is broken and needs to be repaired (Leskovec et al., 2014); second, the most common sequences of web pages visited by users can lead to understand how to best structure a website; finally, the consumer behavior on a web site could help to recommend the products that may interest him (recommender systems). Users of modern e-commerce services are exposed to a myriad of diverse product choices. Helping users to find what they are looking for can mean a huge increase in user satisfaction and company profit. The goal of recommender systems is to analyze data associated with users and products to give personalized recommendations. Recommender systems generally fall into two categories: content based strategies, that involve features directly associated with the users and products (age, sex, demographic features, etc.) and collaborative filtering, that analyses users' past behavior to predict how users will act in the future. Collaborative filterings aim to model relationships betweens users and items using two different kind of data: explicit feedback data such as user-item ratings and implicit feedback data such as clicks, page views, music streaming play counts. Recently, there is an increasing demand for collaborative filtering methods that are designed for implicit data because web-data mainly comes in this form and they can be collected without user's explicit sentiment. In this context, cluster analisys is a popular technique for detecting intrusions and reporting the behaviour of unique users over the time. Every single stream of observations can be seen as a function. Normally, in the data stream, the functional form can vary in time; however, sometimes the theoretical distribution of the data is know in advance. For example, when a new product is launched on the market, at a first stage the product undergoes an increase almost exponentially, while in a second phase the growth diminishes to almost nothing and stabilizes to a certain maximum threshold. In this case, the distribution assumes the classic shape to "s" typical of the logistic curve. That is the"life cycle of a product" composed by the four phases: introduction, growth, maturity and decline. For example, in the data streaming context, we could assume that the products are the goods for sale on a web site, or the hypertests of a web page, or the content of a web page. Data flowing from the server of a web site, such as the number of clicks on hypertests of a web page, or the number of like on different content of a web page, can be imagined as a single continuous streams of data in the time domain. Indeed, in phenomena like data streaming, the observations for each statistical unit can be expressed through a process defined by a curve or a function. However, in this context, a great problem is to compute summary statistics of the same functional form as the observed data (De Sanctis and Di Battista, 2012). This issue exists when the functional data constitute a subset which is not a linear subspace in the functional vector space. Specifically, in this paper we focus on cases where the functional form is known a-priori and we propose a functional classification.

2. Johnson's Probabilistic Approach

In this paper, we focus on probabilistic approaches that consider the probability of an user choosing to interact with an item. In the literature, there is a vast amount of proposals. For example, Goplan et al.(2013) introduced a factorization model that factorizes users and items by the Poisson distributions. Instead, Johnson (2014) proposed a probabilistic framework for the implicit case in which he models the probability of a user choosing an item by the logistic function. He uses a matrix factorization model that is a subset of a larger family of models known as latent factor models. The model focuses on collaborative filtering recommender systems with implicit feedback data. In particular, let be: $U = (u_1, u_2, ..., u_n)$ a group of n users, $I = (i_1, i_2, ..., i_m)$ a group of m items and $R = (r_{ui})_{n \times m}$ an user-item observation matrix where $r_{ui} \in \Re > 0$ represents the number of times that user u interacted with item i. The r_{ui} values are non-negative feedback values and not required to be integers but are any non-negative reals. This is to allow contextual or temporal weighting of observations. For example, we may choose to weight more recent user streams higher than older streams as a users taste may change slightly over time. In most cases $\mathbf{R}$ is a very sparse matrix as most users only interact with a small number of items in I. If user u does not interact with item i we place 0; but this does not necessarily imply that the user does not prefer the item, it could simply imply that the user does not know about the item. In particular, latent factor models attempt to uncover latent or unobserved factors to encode the users and items in U and I. Then, the goal is to find the top recommended items for each user and for each item that they have not yet interacted with. Johnson's probabilistic approach consists in factorizing the observation matrix $\mathbf{R}$ by two lower dimensional matrices $\boldsymbol{X}_{n \times f}$ and $\boldsymbol{Y}_{m \times f}$ where f is the number of latent factors. The rows of $\mathbf{X}$ are latent factor vectors that represent a users taste while the columns of $\boldsymbol{Y}^T$ are latent factor vectors that represent an items style, genre, or other implicit characteristics. Let $l_{u,i}$ denotes the event that user u has chosen to interact with item i (user u prefers item i). Then, let the probability of this event occurring be fitted by a logistic function parameterized by the sum of the inner product of user and item latent factor vectors and user and item biases as follows (Johnson, 2014):

$$p(l_{ui}|x_u, y_i, \beta_u, \beta_i) = \frac{exp(x_u y_i^T + \beta_u + \beta_i)}{1 + exp(x_u y_i^T + \beta_u + \beta_i)} \tag{1}$$

Here, β_u and β_i represent user and item biases which are meant to account for variation in behavior across both users and items. Some users will have a tendency to interact with a diverse assortment of items in I while others will only interact with a small subset. Similarly, some items will be very popular and so will have a high expectation of being interacted with across a broad audience while other items will be less popular. The bias terms are latent factors associated with each user and item that are meant to offset these behavior and popularity biases.

3. The functional approach

Data flowing from the server of a web site, such as the number of clicks on hypertests or the number of like on different content of a web page, can be imagined as a single continuous streams of data in the time domain. Thus, every single stream of observation can be seen as a function. An appropriate approach to deal with this kind of data is the functional data analysis (FDA) approach (Ramsay and Silverman, 2005). Indeed, it addresses problems in which the observations are described by functions rather than finite dimensional vectors. The functional datum should be regarded as a single entity, instead of a sequence of observations; it is a single measure along a continuum. Although the functional datum lives in a continuum, the empirical observation must necessarily refer to the discretization of the domain; thus, in real applications, functional data are often observed as a sequence of point data. The FDA approach is able to reconstruct the functional datum as a whole, starting from raw observations of the reference domain. However, De Sanctis and Di Battista (2012) and Di Battista et al. (2016) highlight that the above approach aims primarily to convert discrete observations to functional form by means of appropriate techniques, such as the use of basis functions. This logic is sustainable when the function underlying the data is unknown in its form and, thus, it is necessary to resort to its approximation. However, there are situations in which the use of smoothing techniques is not suitable. In fact, when the underlying data process is known, the approximation by means of basis functions alters the measure of the observed phenomenon because the latter is inherent in the functional datum.

In particular, statistics derived from the classic functional approach (Ferraty and Vieu, 2006; Ramsay and Silverman, 2005) may not be able to explain the characteristics of the underlying functions and this, obviously, leads to misinterpretations of the same functional statistic. From a statistical point of view, the summary statistics should be expressed in the same unit of measurement as the data, which, in our specific case, is expressed by the observed function. Therefore, in this setting, we aims to find a statistic of the same functional form as the observed data, in order to provide a correct interpretation of the phenomenon under study (Di Battista and Fortuna, 2013). Normally, in data stream, the functional form can vary in time but sometimes, the theoretical distribution of the data is know in advance, such as in the Johnson's approach (see Section 2). This paper focuses on this particular aspect of functional data analysis. In more detail, we obtain a functional mean of the same functional form of the data. In order to reach this aim, we work on subsets of a Banach space, which are convex, or for which there exists a transformation one to one to a convex (Di Battista et al., 2015).

4. Functional statistics on functional spaces

Let X be an arbitrary measure space with a positive measure μ. $L^p(\mu)$, $p > 0$, denotes the set of all real or complex measurable functions f on X for which:

$$\int_X |f(x)|^p d\mu(x) < \infty \tag{2}$$

(Rudin, 1986). It becomes a real or complex vector space because $L^p(\mu)$ is closed under its natural operations: whenever $f, g \in L^p(\mu)$, then $f + g \in L^p(\mu)$ and whenever $f \in L^p(\mu)$ and α is a scalar, then $\alpha f \in L^p(\mu)$.

Moreover $L^p(\mu)$ is a normated space with respect to the following norm:

$$||f||_p = \left(\int_X |f(x)|^p d\mu(x) \right)^{\frac{1}{p}} \qquad \forall f \in L^P(\mu) \tag{3}$$

Therefore it is also a metric space with respect to the distance induced by the norm:

$$d\Big(f,g\Big) = ||f-g||_{L^p} = \left(\int_X |f(x) - g(x)|^p d\mu(x) \right)^{\frac{1}{p}} \tag{4}$$

Finally, $L^p(\mu)$ is a complete metric space with respect to the metric derived from its norm, that is, every Cauchy sequence in $L^p(\mu)$ converges to an element of $L^p(\mu)$. This means that it is a Banach space. We recall that a subset S of a generic vector space V is called a subspace if it is itself a vector space. We note also that a subset S of a vector space V is said to be convex if, for each point, $c, v \in S$, the point z defined as $z = (1-t)c+tv$ is an element of S, $\forall t \in [0,1]$. In our setting, the functional data observed for each unit $f_i(x)$, $i = 1, 2, ..., n$, belong to $L^p(X)$, $p > 0$, which is the functional space of the real functions, defined on $X \subset \mathbb{R}^k$, and measurable, with respect to the euclidean measure, which satisfy equation (2). As stated before, it is a Banach space with the norm defined in equation (3). We are interested in convex subsets of $L^p(X)$. Let S be a convex subset of $L^p(X)$. For induction, we can easy prove that, if $f_1, f_2, ..., f_n$ are elements of S, then their functional mean, defined by $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ is an element of S. In order to define the functional mean on a generic non convex subset, we assume the following (Di Battista et al., 2015):

HYPOTHESIS: There is a biunivocal transformation, say T, between a non-convex subset S and a convex subset S' in $L^p(X)$. If there exists such a transformation T, it is possible to consider the functional mean, as the value of synthesis, in the convex subset S'. Then we can return it to S through the inverse transformation T^{-1}. Under this hypothesis, the following definition can be stated:

Definition 1 *Let $f_1, f_2, ..., f_n$ be functions in S and $g_1, g_2, ..., g_n$ the corresponding functions in S' defined by $T(f_i) = g_i$, $i = 1, 2, ..., n$. If $g = \frac{1}{n}\sum_{i=1}^{n} g_i$ is the functional mean in S', the only element f in S, which corresponds to g by T^{-1}, that is $f = T^{-1}g$, is defined to be the functional mean in S.*

In conclusion, we propose a functional mean on a subset S of a Banach space in the following cases: when S is convex we can define the functional mean in the usual way; when S is not convex, we are often able to find a transformation one to one, T, from S to a convex subset S', thus, we define the usual mean of the transformed data in the convex subset S' and we return to S through the inverse transformation T^{-1}. In both cases it is possible to obtain functional statistics of the same functional form of the data.

5. Classification on functional space

Clustering of functional data is generally a difficult task because units belong to an infinite dimensional space where the equivalence between norms and distances, which is typical in a finite dimensional euclidean space, fails (Ferraty and Vieu, 2006). For this reason, different clustering approaches have been proposed over the years. The most simple one is called raw-data clustering and consists of using direct discretization of the functions at some time points. Obviously, this procedure presents many limitations since it ignores the functional nature of the observations. Indeed the most commonly used approaches are mainly based on three methodologies: dimension reduction before clustering, nonparametric methods and model based clustering methods. The first one (called two-stage method) approximates functional data with elements from some finite dimensional space such as coefficients of functional data expansion or a given number of principal components and then applies classical clustering algorithms (Peng and Muller, 2008). Nonparametric methods for clustering, instead, consist generally in defining specific distances or dissimilarities between curves and then apply clustering algorithms for finite dimensional data. Finally, model based clustering methods assume a probabilistic distribution underlying the data on either the principal components or basis expansion coefficients. The disadvantage of the above methods is that clustering results can differ depending on how the curves fit to the data. In particular, estimating curves using different sets of basis functions corresponds to different linear combinations of the data and many clustering algorithms, such as K-means, are not invariant to linear transformations of the data (Tarpey, 2007). In this paper we focus on a nonparametric clustering method applying the L^p distance in (4) and a k-means algorithm (Forgy, 1965). In order to reach this aim, at first, we select $p = 2$. Indeed we recall that $L^2(\mu)$ is, among all $L^p(\mu)$ spaces, the only Hilbert space, that is its norm is induced by the inner product:

$$(f, g) = \int_X f(x)g(x)d\mu(x) \tag{5}$$

whenever $f, g \in L^2(\mu)$. This choice of p is the best in order to solve consequent problems of minimality of the distance. Then, if S is a convex subset of $L^2(X)$, we define an iterative procedure which is initialized by fixing the number k of clusters $\{C_1, C_2, ..., C_k\}$ and by selecting k arbitrary and distinct initial centroids $\{\phi_1^0, ..., \phi_k^0\}$, one for each cluster. At the m-th algorithm iteration, $m > 0$, each function f_i, $i = 1, 2, ..., n$, belonging to a given functional data set $\{f_1, f_2, ..., f_n\}$, is assigned to the cluster whose centroid, at the $(m-1)$-th iteration, is the nearest according to the chosen distance:

$$\arg\min_{l=1,2,..,k} \left(\int_X |f_i(x) - \phi_l^{m-1}(x)|^2 dx \right)^{\frac{1}{2}} \tag{6}$$

Once all of the functions have been assigned to a cluster, a new centroid ϕ_l^m in the l-th cluster C_l is recalculated as functional mean of the data set's functions belonging to it:

$$\phi_l^m = \sum_{f_i \in C_l} \frac{f_i}{n_l} \tag{7}$$

where n_l is the number of functions in C_l. This procedure continues until no function changes cluster. We remark that the algorithm is possible because S is a convex subset, then, by recalculating new centroids, they remain in S. If S is not convex but it is known a transformation T such that $S' = T(S)$ is convex, all the procedure can be done in S' then transported on S by the inverse transformation T^{-1}. In this paper, we aim to apply this method to data stream analysis; in particular, in order to identify specific common patterns, a functional k-means algorithm can be implemented, specifying an L^2 metric directly on the explicit known form of the function in equation (1). To obtain a functional mean belonging to the same space of the initial functions, we use a transformation T between a non convex subset and a convex subset in $L^p(X)$. In particular, T can be expressed as follows:

$$T\Big(p(l_{ui}|x_u, y_i, \beta_u, \beta_i))\Big) = logit\Big(p(l_{ui}|x_u, y_i, \beta_u, \beta_i)\Big) = x_u y_i^T + \beta_u + \beta_i \tag{8}$$

Thus, the classification can be implemented on the obtained linear functions. After data have been clustered, we apply the inverse transformation of the log odds to come back the same functional form of the observed data. In particular the combined use of parametric FDA and functional classification methods, allows us to identify different patterns of probability that the users interact with the items on a web site. In many phenomena the observations for each statistical unit are expressed through a process defined by a curve or a function. Because the characteristics of data stream are that data continuously flow, data size is extremely large and potentially infinite, functional data analysis is a suitable approach. In data stream context, normally the functional form can vary; however, sometimes the theoretical distribution of the data is known in advance such as in the Johnson's model. In these cases it is possible to apply our method to classify logistic functions.

References

Cobb C, Douglas P. (1928). A theory of production. American Economic Review 18 (1):139-165.

De Sanctis A., Di Battista T. (2012). Functional analysis for parametric families of functional data. International Journal of Bifurcation and Chaos 22 (9):1250,22611250,2266.

Di Battista T., Fortuna F. (2013). Assessing biodiversity profile through fda.Statistica 1:69-85.

Di Battista T., Fortuna F., Maturo F. (2016). Environmental monitoring through functional biodiversity tools. Ecological Indicators 60:237-247.

Di Battista T., Fortuna F., De Sanctis A. (2015). Functional statistics on functional spaces. ADAC (Under Review).

Ferraty, F., Vieu, P. (2006). Nonparametric functional data analysis. Springer, New York.

Forgy E. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21:768-769.

Gattone SA, Di Battista T (2009). A functional approach to diversity profiles. Journal of the Royal Statistical Society 58:267-284.

Goplan P. & Hofman J. & Blei D. (2013). Scalable Recommendation with Poisson Factorization arXiv preprint arXiv 1311-1704.

Johnson C., (2014). Logistic Matrix Factorization for Implicit Feedback Data.

Leskovec J., Rajaraman A., Ullman J.D. (2014). Mining of Massive Datasets, Cambridge University Press.

Masud M.M., Woolam C., Gao J., Khan L., Han J., Hamlen K.W., Oza N.C., (2011). Facing the reality of data stream classification: coping with scarcity of labeled data, Springer-Verlag London Limited.

O'Callaghan L., Mishra N., Meyerson A., Guha S., Motwani R. (2002). Streaming-data algorithms for high-quality clustering. In Proceedings of the 18th International Conference on Data Engineering. IEEE.

Parikh D., Tirkha P., (2008). In Data Streams Using Classification And Clustering Different Techniques To Find Novel Class, International Journal of Research in Engineering and Technology, p163, vol.2.

Patil G.P., Taillie C. (1982). Diversity as a concept and its measurement. Journal of the American Statistical Association 77:548-567.

Peng J., Muller H. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. The Annals of Applied Statistics 2(3):1056-1077.

Ramsay, J. and Silverman, B. (2005). Functional Data Analysis, 2nd edn. Springer, New York.

Ratcliffe S.J., Leader L.R., Heller G.Z. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. ii: functional logistic regression. Stat Med 21 (8):1115-1127.

Rossi N, Wang X, Ramsay J (2002). Nonparametric item response function estimates with the em algorithm. Journal of Educational and Behavioral Statistics 27:291-317.

Rudin W. (1986). Real and complex analysis. McGraw-Hill.

Tarpey T (2007). Linear transformations and the k-means clustering algorithm: applications to clustering curves. The American Statistician 61(1):34-40.

Vieria S., Hoffmann R. (1977). Comparison of the logistic and the gompertz growth functions considering additive and multiplicative error terms. Applied Statistics 26:143-148.