Knowledge Extraction by Dynamical Clustering of sea waves streaming data

Elvira Romano[1], Antonio Balzanella[1], Rosanna Verde[2]

[1] Dipartimento di Matematica e Statistica Universitá degli Studi di Napoli "Federico II", Via Cintia Complesso Monte Sant'Angelo I-80126, Napoli
[2] Dipartimento di Studi Europei e Mediterranei, Seconda Universitá di Napoli, San Leucio

Abstract. Data stream can be thought as a sequence of ordered data items, where the input arrives more or less continuously as time progress. There exist several applications producing data stream, e.g. telecommunication system, stock markets customer click streams etc.. In this paper we consider the problem of extracting knowledge by a dinamical clustering algorithm of sea waves streaming data, that is to say evolving streaming of data coming from a multisensor system. For this purpose we develop an updating version of Dynamical Clustering Algorithm [5]. This problem is very interesting from a practical point of view. It is based on the computation of a prototypal wave through a free-knot smoothing spline, optimizing a non linear problem. Thanks to this approach, it is possible to investigate in which way the incoming data change according to the various steps of process registration, and to have a summary description of the entire data thought prototypals flowing curves using a small amount of memory and time.

Keywords: data stream, data mining, clustering, sea waves propagation, free knots spline.

1 Introduction

In applications such as network monitoring, telecommunications data management, clickstream monitoring, manufacturing, sensor networks, and others, data takes the form of continuous data streams rather than finite stored data sets. These data are created by continuous activity over long periods of time and are therefore data which grow continuously over time. This has lead to the development of new strategies able to furnish an overview of the key characteristics in the data which change over time in a fast way. In the last years since it has been advances in hardware technology the analysis of such kind of data has gained more attention. An important area of interest is that of clustering. The clustering problem is especially interesting because of its application to data summarization and outlier detection, but is a difficult problem to solve because of large volumes of data arriving in a stream lead most traditional algorithms too inefficient. Regarding aspects of data storage, management and processing, the continuous arrival of data items in multiple, rapid, time-varying requires clustering adaptive strategies in the sense to up-todate clusters, and able to taking new data items into consideration as soon as they arrive. The clustering problem has addressed many scientific area such as the database, data mining and statistics communities [3], [6],[10]. Despite these work presents interesting methods to classify datastream they have the aim to classify elements of one individual data stream, which is quite different from the problem that we consider here. In this paper we are going to analyze the streams themselves rather than single data items thereof. Our approach is different on the way of performing clustering. It share some ideas with the field of Dynamical Curve Clustering [14], [13], where the main aim is to classify a set of time stamped curves. We develop an updating version of Dynamical curve clustering where our focus is the analysis of *stream of curves* coming from an applicative study on sea waves streaming data. The idea is to classify this data using sliding windows of fixed size. The method incorporates good quality of clusters combined with efficient computational properties. It is incrementally updatable and is highly scalable on both the number of dimensions and the size of the data streams.

The remainder of the paper is organized as follows: section 2 provides some background information on clustering datastream; the main topics of our strategy are proposed in section 3; in section 4 main results deriving from the applicative field are discussed.

2 Overview on Clustering DataStream

The online nature of data streams and their potentially high arrival rates imposes high requirements on clustering [1]. Several adaptations of standard statistical and data analysis methods to data stream have been developed recently. It is known that solving a clustering problem is equivalent to find the global optimal solution of a non-linear optimization problem, forcing the need to apply optimization heuristics [12].
Although extended clustering on datastream are generally blind to the evolution of the data [2][10], they do not address some important issues such as cluster quality and characterization. In these works however, as stayed before, the problem is to cluster the elements of a single stream, which is clearly different from our problem, where the objects to be clustered are streaming curves or time series rather than single data itemes. The parallel explosions of interest in streaming data, and data mining of time series have had surprisingly little intersection. This is in spite of the fact that time series data are typically streaming data [9]. In literature this topic has been handled in several ways, the two major approaches are based on: a symbolic representation of time streaming data from one side and on incrementally constructs algorithms on the other. The first one consists in transforming real valued time series into symbolic representations,[8], [9]. However fatal flaws in such representation exist. Firstly the dimensionality of the symbolic representation is the same as the original data, and virtually all data mining algorithms scale poorly with dimensionality.
Moreover, before creating the symbolic representation,most of these symbolic approaches require one to have access to all the data. This last feature explicitly thwarts efforts to use the representations with streaming algorithms.
The second one consists in dealing with time series streaming through particular up-to-date functions avoiding symbolic assumption for their description.One of these is the Online Divisive-Agglomerative Clustering (ODAC) system. It furnishes an incremental implementation of divisive analysis clustering, using the correlation between time series as similarity measure [12]. Although it is very interesting from a computational point of view and the system is able to exhibit dynamic behavior, adapting to changes in time-series, a drawback of the method as observed by the authors is that the agglomeration phase does not consider for changes in the cluster structure.
Apart from its practical relevance, the description of cluster structure such as also indexes for clustering evaluation play important role in clustering methods. In this sense our goal is to develop a strategy that meet these requirements. As we shall demonstrate our procedure contrary to the existent methods tries to take into account both, the representation of the cluster and the best optimization problem for computational point of view.

3 The Updating Dynamical clustering strategy

In the previous section we have shortly presented the problem of clustering data stream and where our method takes place. In this section we propose the Dynamical clustering strategy and its more salient characteristics.
When considering curve streams data in a sliding window of length T, a curve stream y can formally be written as a T-dimensional vector $y = (y_1, \ldots, y_T)$, where a single observation can be considered as a curve. We aim to determinate a partition of our datastream into C cluster such that each of them is characterized through a best representative curve, called prototype.
This type of approach comes from the applicative field of sea waves profiles, each single entities or sea waves streaming curve derives from several independent multisensor systems located in several submerged places. We propose to classify these streaming curves according an updating version of DCA performed on sliding windows. Each windows represents a time stamped interval of a

streaming curve. The classification strategy is performed in an recursive way such that the sets of prototypes is updated according to the changes of the sliding windows. Especially an optimal criteria to capture the maximum information with the minimum computational complexity is proposed. An important characteristic of the method consists in performing the clustering strategy with free knots spline estimation of the prototypal curve. A free knot spline is a spline where the knot locations are considered parameters to be estimated from the data. It means a non parametric estimate of a representative curve that allows to take into account the similar slope changes of all the curves in the cluster. The representation of the information depends on the location and the number of the knots. Choosing a set of knots, is a more controversial issue since the same set of knots is not universally good. So an iterative procedure that simultaneously computes a free-knot spline estimators of the prototypes of the clusters and is able to classify the curves in homogeneous classes is proposed for each sliding window. According to the optimized criterion, starting from an initialization step we obtain, for each one of the C clusters, a local model prototype identified from the best set of knots. Each curves is assigned to a class according to its proximity to the prototype in the sense of mean square error. After the initial allocation, we compute a weights-vector W starting from the clusters, where each element is computed considering the number of curves fresh-assigned to the cluster. With the arrival of a new stream of data (in the sense of a new matrix coming from a sliding window), we compute a new set of C prototypes still identified from the best set of knots, taking into account the new curves and the old set of prototypes. An important feature of the clustering method consists in obtaining cluster's characterization through *a streaming curve which describes the prototypal profile of the cluster* at each iteration. It introduces the advantage to not only gather the course but also the shape of every flowing curve. Formally for each sliding windows the core of the methodology is to optimize the following criterion:

$$\Delta(P,G) = \sum_{c=1}^{C} \sum_{i \in P_c} \mu_c \delta^2(y_j^i, g_c) \;\; P_c \in P, g_c \in G \tag{1}$$

where $\mu_c = \frac{1}{|P_c|}$ is a weight, $\delta^2(y_j^i, g_c) = \left\| y_j^i - g_c \right\|_2$ is the L_2 distance and the system of $G = \{g_1, \ldots, g_C\}$ of class prototype are computed by optimizing an adequacy criterion
$\phi(g) = \sum_{i \in P_c} \delta^2(y_j^i, g_c)$ that leads to a free-knot spline estimators of the representative curve prototype.
More precisely, let $\xi^c = \{\xi_1^c, \ldots, \xi_M^c\}$ $\forall c \in P_c$ be a vector of knot sequence, we look for a good approximation for the prototypal discretized curve $g_c(\xi^c) \subset R^{n_i}$, that is the estimate a representative function of a set of curves, solution of a least squares problem. It is an approximation of the barycenter of a set of curves in the space $S_{H,M} = \bigcup_{\xi^c} S_{H,M}(\xi^c)$, of polynomial spline order $H \geq 1$ with knot sequence $\xi^c \in \Re^M$ such that it does not resort an individual smoothing. For each of this prototoypal discretized function, since any function of $S_{H,M}$ can be writed as:

$$g_c(\alpha^c, \xi^c) = \sum_{l=1}^{M+H} B_{l,h}(\xi^c) \alpha_m^c \tag{2}$$

where B_{l,h,ξ^c} denote the usual B-spline basis function of order H with a sequence $\{\xi^c\}_{m=1}^M$ of knots and α_m^c are the sequence of B-spline coefficients. According to the optimized criterion, initializing the procedure with equispaced knots we obtain for each cluster a local model prototype identified from the best sets of knots.
Each curves is assigned to a class according its proximity to the prototype in the sense of mean square errors. The true functional form of each function prototype $g_c(\mathbf{t}) = (g_c(t_1), \ldots, g_c(t_J))^t$ is a function of the generic variable t with an optimum vector of knots $\hat{\zeta}^c$. The dimension of the vector of the knots which characterizes this function is such that to minimize the loss of variety of *true functions*, where the loss is defined to be the square root of mean of squared distance from the truth to the estimated prototypal function at the design points. Moreover each cluster can be described as j-dimensional vector of curves that follows the model:

$$\mathbf{y^i} = \mathbf{B}(\zeta^\mathbf{c})\alpha_1 + c^\mathbf{i} \quad \forall \mathbf{i} \in \quad \mathbf{C} \tag{3}$$

where ϵ^i is an observation noise with mean zero, constant variance and covariance zero for distinct arguments values and B is the matrix of B-spline function [13]. In this sense these are the *local model prototypes* able to describe the several obtained clusters.
Since the dependence of the basis functions from the knots is nonlinear, the optimized problem is a nonlinear minimization problem. Despite it could seems an intensive computational problem we"ll demonstrate how it may be developed in an efficient way. The solution to this problem is extensively explained in [13]. It consists in a little modification of the algorithm proposed by Gervini [7]. This algorithm produces a set prototypes related o several sequences of vectors of knots, among these the minimizers of the Generalized Cross Validation Criteria (GCV) are chosen. Let's see a general description of the updating algorithm.

4 The Updating Dynamical clustering algorithm

In this section we discuss algorithm we propose. It consists in two phases, an initialization phase and an updating phase. These are respectively performed as follows:

*Initialization phase*where these steps are performed:

- the optimal set of prototypes g_c taking into account the data coming from the first sliding window *sw* are computed,
- the curves are assigned to the cluster according the Mean square error,
- the number of curves n_c in each cluster are stored in a weight vector ,
- representative curves of each cluster, called microprototypes, are identified. The criteria proposed to identify these microprototypes are several.They depends from the variability of each cluster. These microprotypes are able to obtain a more detailed information on the cluster structure, that can be summarized by a quantification of the number of the curves more similar to the identified microprototypes (in the sense of mean square error).

*Updating phase*where for each *sw* the prototype and classes of curves are updated following this scheme:

- the prototype are computed taking into consideration both the stream of curves in the current *sw* and the prototypes of the old sliding window weighted by n_c;
  allocation step: the streaming curves of the current sliding window are allocated to the clusters according to the criterion of Euclidean distance or some variation of it;
- the number of curves n_c in each cluster is updated;
- for each microprotype the nearest prototype in the sense of mean square error is identified. If this microprototype is not the nearest to the original prototype (from which it takes place), then the weights are updated adding the number of the elements that quantify the microprotypes to the nearest prototype and subtracting the same quantity to the prototype of the cluster which the microprototype belongs to;
- a new set of microprototypes are identified, recording the new number of curves more similar to these last ones.

5 Sea waves streaming Analysis

In the view of understanding the wave breaking process in presence of submerged trapezoidal porous barriers several experimental analysis were conducted from the department of Hydraulic and Environmental Engineering "Girolamo Ippolito", in wave flume with glass walls; rubble mound breakwaters with varying geometry and permeability were tested. The engineering problem usually consists in analyzing the role of the main breaker types, that takes place at submerged breakwaters, in determining the characteristics of the wave profile in protected area. Since the detection of slope change point depends from the breaker type, several structure are considered. For a each

of them a multysensor system reveals the rupture of each sea waves on the several breakwaters. In this context, the data produced by sensors, assume the form of flows of curves evolving continuously over time so they constitute a suitable platform to evaluate our methodology. They consist of 10000 streaming curves of size 100 from several multysensor system located on different place. The aim is to furnish, analyzing the sea waves streaming curves, classes of streaming curves and a set of prototypes representative of each cluster able to provide an helpful summary from a physical point of view. Our strategy on streaming curves has been implemented on data coming from several real experiments on sea waves in presence of submerged trapezoidal porous barriers, and then compared to Dynamical curves clustering with free knots spline Estimation([13]) on batched data. The results have been encouraging both from the classification point of view (in terms of within homogeneity of the clusters and separation among them) and from the ability of the prototypes to characterize clusters. We can deem that the proposed method, represents a good compromise between computational requirements and goodness of outcomes. It can be confirmed from the graphical representation that we obtain Figure 1. Each of them underlines different shapes which

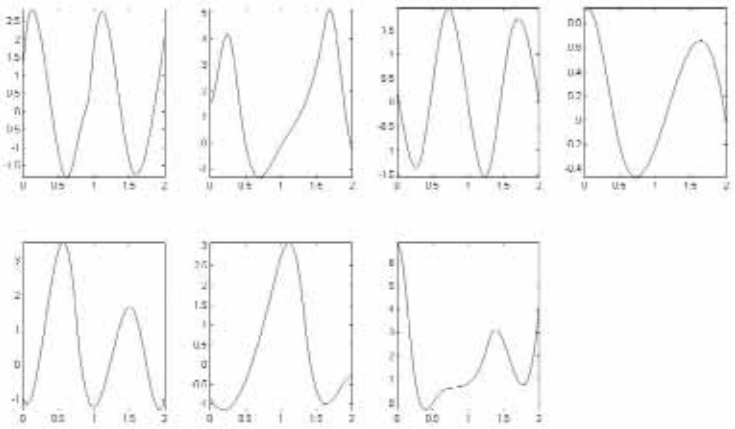

Fig. 1. Prototypal curves

represent characteristic behavior of sea waves streaming curves. An other measure of the clustering quality is furnished from the quality partition index Q defined as the ratio between the within and the total variability of each cluster of streaming curves. They ranchs around 0.5.

6 Conclusions and Perspectives

Overall, the proposed method points to a good trade off between a statical and a computational approach. In several applications problem, speed computation is the principal objective, however mistakes can be costly. Our Updating Dynamical clustering algorithm, on the contrary, exhibits superior performance in the perspective to catch meaningful characteristic of the cluster structure. Actually we are working on an improving version of the proposed method from computational point of view. Object of further research will be the comparison with other possible methods on such kind of data.

References

1. Aggarwal C. C.,Han, J., Wang, J., Yu, P.S.: On demand classification of data streams, On Demand Classification of Data Streams, Proc. 2004 Int. Conf. on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA, August (2004).
2. Beringer, J., Hullermeier, E.: Online Clustering of Parallel Data Streams, Data Knowledge Engineering, (2005).
3. Bradley, P. Fayyad, U., Reina, C.: Scaling Clustering Algorithms to Large Databases. SIGKDD Conference, (1998).
4. Cao, F., Ester, M., Qian, W. and Zhou, A.: Density-based Clustering over an Evolving Data Stream with Noise, To appear in Proceedings of the 2006 SIAM Conference on Data Mining (SDM'2006).
5. Diday, E.: La Method des nuées dinamiques, Rev. Stat.Appliques. XXX 2, 19–34, (1971).
6. Domingos, P. Hulten, G. : Mining High-Speed Data Streams. ACM SIGKDD Conference, (2000).
7. Gervini, D.: Free-knot spline smoothing for functional data, to appear in Journal of the Royal Statistical Society,(Series B), (2006).
8. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An Online Algorithm for Segmenting Time Series. In Proceedings of IEEE International Conference on Data Mining. 289–296 ,(2001).
9. Lin, J., Keogh, E., Lonardi, S. and Chiu, B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13, (2003).
10. O'Callaghan, L. et al. Streaming-Data Algorithms For High-Quality Clustering. ICDE Conference, (2002).
11. Ordonez, C.: Clustering Binary Data Streams with K-means, ACM DMKD (2003).
12. Rodrigues, P., Gama, J., Pedroso J. P.: Hierarchical Time-Series Clustering for Data Streams, Proceedings of First International Workshop on Knowledge Discovery in Data Streams, 24 September, Pisa, Italy, (2004).
13. Romano E.: Dynamical curves clustering with free knots spline estimation. Methodological contributions and Applications. Phd Thesis, 30 November (2006).
14. Romano E., Verde R., Lechevallier Y.: Dynamical classification of functional data with free knots spline estimation, Proceedings of Knowledge Extraction and Modelling, IASC-INTERFACE-IFCS Workshop, 4-6 Capri, Italy, (2006).