Knowledge Extraction by Dynamical Clustering of sea waves streaming data

Elvira Romano* Antonio Balzanella* Rosanna Verde**

* Università degli Studi di Napoli Federico II
Via Cintia, Complesso Monte Sant' Antangelo
I-80126 Napoli
elvroman@unina.it
balzanella2@alice.it

** Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea "Jean Monnet", Seconda Università degli Studi di Napoli
Caserta, I-81020, Italy
rosanna.verde@unina2.it

Aims

In the present paper we aim to define a strategy able to extract in *accurate* way information from sea waves streaming data.

The problem derives from an applicative field:
evolving streaming data coming from multisensor system .

The proposal:

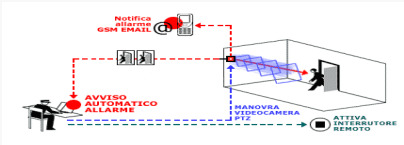
An extension of Dynamical curves clustering with free knots spline estimation

The sensitivity of the proposed method is investigated using a dataset coming from an experimental study conducted by the Department of Hydraulic and Environmental Engineering Girolamo Ippolito of Naples*.

*Panisi, F., Calabrese, M., Buccino M. (2006). Breaker Types and Free Waves Generation at Submerged Breakwaters in Proceeding of XXIX of HYDRA'06, Roma.

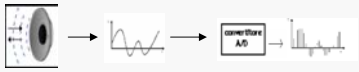
Other possible applicative fields:

Video Surveillance



Scentific Researches

Vocal recognition



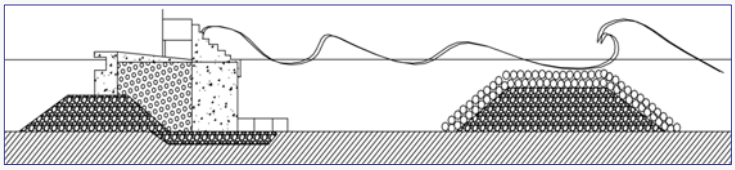
Roadmap

- ✓ Sea waves streaming data structure
- ✓ An overview on clustering of streaming curves
- ✓ Our proposal & The main aspects of the proposed strategy
- ✓ The Updating Algorithm
- ✓The optimization problem & Cluster characterization
- ✓Case study: Streaming curves data, propagation of sea waves
- ✓Conclusion and Prospectives

Sea waves streaming data structure

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.

A multysensor system reveals the rupture of each sea waves on the several breakwaters. The data produced by sensors, assume the form of flows of curves evolving continuously over time.



Overview on clustering streaming curves

In literature clustering curve datastream is usually known as time series streameanig data clustering. It has been handled in several ways, the two major approaches are based on:

Symbolic representations

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).

Incrementally construct

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).

Drawbacks:

The dimensionality of the symbolic representation is the same as the original data, and virtually all data mining algorithms scale poorly with dimensionality.

These are very interesting from a computational point of view and the system is able to exhibit dynamic behavior, adapting to changes in time-series, but usually the agglomeration phase does not consider for changes in the cluster structure.

Our proposal

Consists in furnishing a summary of the key characteristics of the data which change over time in a fast way.

The strategy, we propose, is an updating version of Dynamical Curve Clustering using free knots smoothing spline, performed on sliding windows of fixed size.

DSCA

Principles of the DSCA

- **Representation by *prototype* (cluster model)**
- **Best fitting measure between streaming curves and prototype**

The Updating Algorithm

Initialization phase:

– 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. These are identified as the empirical maximum and minimum values of the obtained prototypal model in a cluster,
– for each cluster, the curves are assigned to their microprototypes according the Mean square error.

Updating phase:

for each sw
- – the prototype are computed taking into account the stream of curves in the current sw starting from the prototypes of the old sliding window (after a transformation in the time axis);
- – 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;
- – a new set of microprototypes are identified, recording the number of curves more similar to these last ones.

Knowledge extraction from phase:

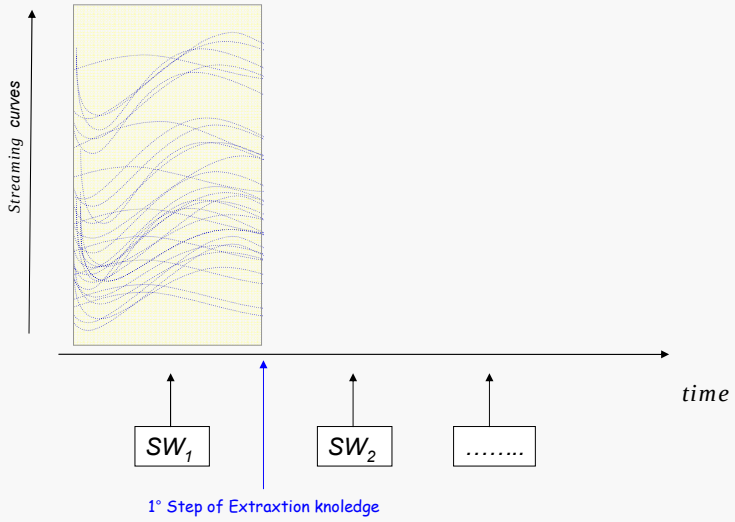
For each sliding window, a set of prototypes and the number of curves belonging to each cluster is outputed

For each sliding window, a set of microprototypes and the number of curves allocated to each one is outputed as further information on the data structure

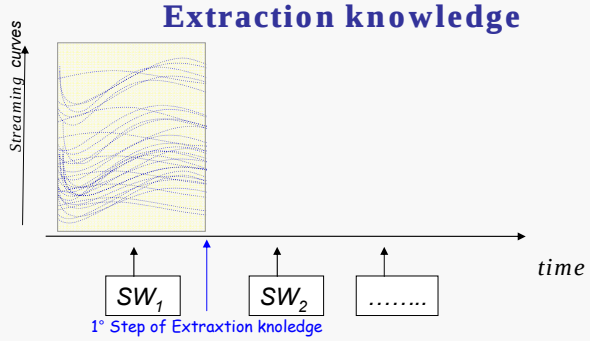
The quality partition index have is able to quantify the goodness of partition and is such to evaluate if there are anomalies in the cluster structure.

Principles of the DSCA

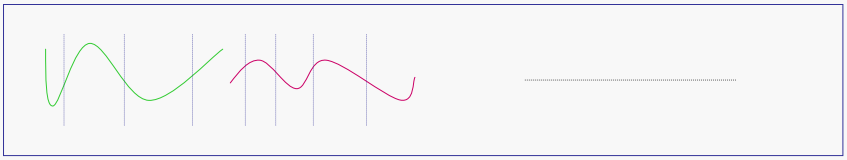
Our proposal



Extraction knowledge



Information Stored: prototypes and best set of knots for each cluster



Note: prototypes are elements representative of each cluster, contrary to other methods where the center is not necessarily representative.

The core of the method: The Dynamical Curves Clustering Algorithm

Dynamical clustering algorithm

It is able to find a way to perform classification based on the *best fitting between the representation function* of the cluster and *the allocation function* to different clusters.

Each prototype is unequivocally determinated by a non linear minimization problem, due to the non linear dependence of the basis curves on the knots.

The advantage to update this method consists in the possibilities to work on the representative elements of the cluster with a dimensionality reduction of the amount of data.

Dynamic clustering algorithm (DCA)

» nuées dynamiques « (Diday, 1972)

Dynamic clustering algorithm optimizes a criterion Δ of
the best fitting between
a partition C of a set Ω of objects w in K classes and
the way to represent the classes $\{C_1,\ldots,C_k,\ldots,C_K\} \in C$

$$\Delta(\hat{C},\hat{G}) = Min\{\Delta(C,G) \mid C \in C_K, G \in G_K\}$$

where : C_K is the set of partitions of Ω in K classes and G_K is the set of the elements representing the K clusters of the partition C_K.

Curve data stream and ANALYSIS

Considering curve streams data in a sliding window of length T,
A curve stream y can formally
be written as a T-dimensional vector $y = (y^1, \ldots, y^T)$

The core of the methodologyfor each sliding window consists in optimizing the following criterion:

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

Where $\mu_c = \frac{1}{|P_c|}$ are weights and $\delta^2(y_j^i, g_c) = \left\| y_j^i - g_c \right\|_2$ is the distance function.

The system of G = {g1, . . . , gC} 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.

The Function prototype and the optimized criterion

Let ξ^c , c C a vector of knots , the aim is to obtain for each cluster a function

$$g_c(\xi^c) \in S_{H,M}(\xi^c)$$

The space of polynomial of order H ≥ 1 with M free knot

$$\phi(\mathbf{y}^i, \mathbf{g}_c) = \sum_{i \in P_c} \delta^2(\mathbf{y}^i, \mathbf{g}_c) = \sum_{i \in P_c} \| \mathbf{y}^i - \mathbf{B}(\xi^c)\alpha_l^c \|^2$$

?

$$\min_{(\xi^c, \alpha_l^c)} \sum_{i \in P_c} \| \mathbf{y}^i - \mathbf{B}(\xi^c)\alpha_l^c \|^2$$

This is a nonlinear minimization problem, being the function prototype dependent from the knots not linearly

where B is the matrix of B-splines basis function of order H with a sequence of knots $\xi^c \in R^M$

How we solve this problem? Firstly....

Jupp transformation of knots vector

$$\varsigma^c = J(\xi^c) = \log \frac{\xi^c_{m+1} - \xi^c_m}{\xi^c_{M+1} - \xi^c_0} \quad m = 1, \ldots, M$$

$$\xi^c_0 = a \qquad \xi^c_{M+1} = b, \quad T = [a, b]$$

$$B_{JxH+M}(\zeta^c) \quad , \quad \xi^c_m < \xi^c_{m+1}$$

The problem become....

$$\phi(\mathbf{y}^i, \mathbf{g}_c) = \sum_{i \in P_c} \delta^2(\mathbf{y}^i, \mathbf{g}_c)_{\alpha_l^c, \varsigma^c} = \sum_{i \in P_c} \| \mathbf{y}^i - \mathbf{B}(\varsigma^c)\alpha_l^c \|^2$$

$$\mathbf{g}_c \xrightarrow{(\hat{\varsigma}^c, \hat{\alpha}^c)} \min_{\varsigma^c, \alpha^c} \sum_{i \in P_c} \| \mathbf{y}^i - \mathbf{B}(\varsigma^c)\alpha_l^c \|^2$$

$$\hat{\mathbf{g}}_c = \mathbf{B}(\varsigma^c)\alpha_l^c$$

optimal

$$for \quad fixed \quad \varsigma^c \quad \alpha_l^c = \left\{\mathbf{B}(\varsigma^c)^T \mathbf{B}(\varsigma^c)\right\}^{-1} \mathbf{B}(\varsigma^c)\bar{\mathbf{y}}$$

$$g_c \rightarrow min_{\varsigma^c} \sum_{i \in P_c} \left\| y^i - B(\varsigma^c) \left\{ B(\varsigma^c)^T B(\varsigma^c) \right\}^{-1} B(\varsigma^c) \bar{y} \right\|^2$$

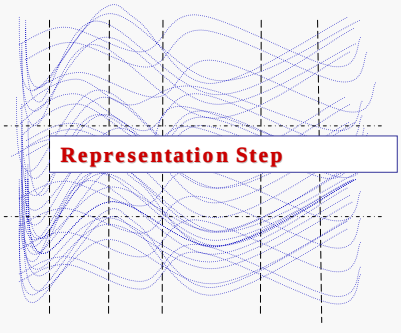
The algorithm to solve this problem is a modified version of Gervini Algorithm. It produces a sequences of knot vectors and a relative function which is optimum, in the sense of mean square error, more precisely It is such to minimize the Generalized Cross Validation Criteria (GCV).

$$CGV(\hat{g}_c) = \frac{\sum_{i \in P_c} \|\mathbf{y^i} - \hat{\boldsymbol{g}}_c\|^2}{J n_c (1 - 2M + H/n_c)^2} \quad i \in P_c$$

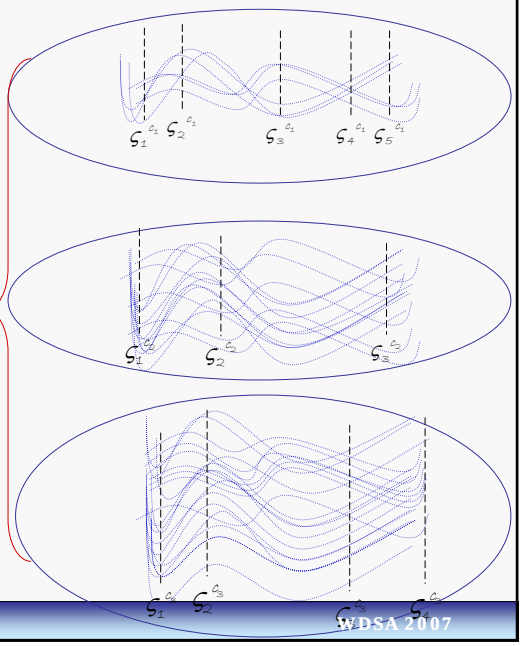
In this way we estimate the 'protoypal function' according to the 'best set of knots'.

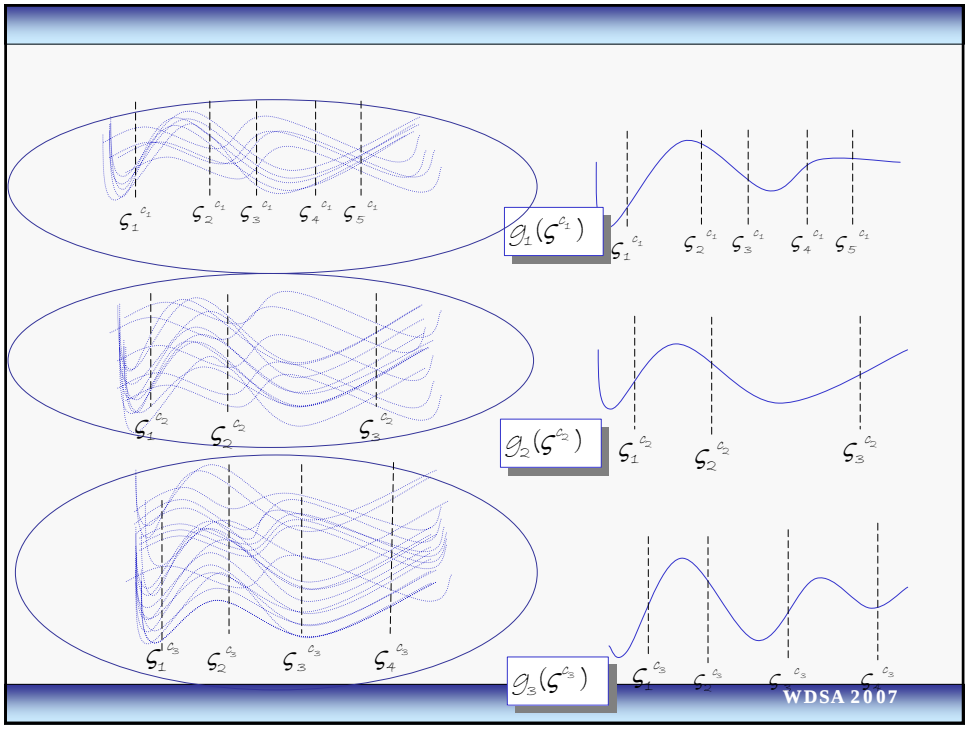
$$g^*_k(\alpha^k, \tau^k) = \sum_{l=1}^{H} B_{l,r,\tau^k}(\tau^{k*}) \alpha^{k*}$$

Dataset for each sliding window



Representation Step





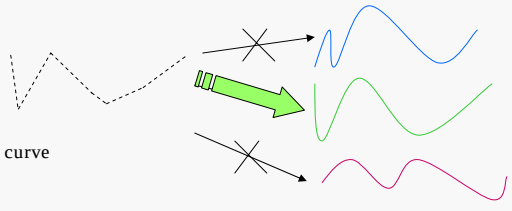
Allocation Step

Allocation Function

a: $G_c \rightarrow P_c$ a(G)=P

with $C_k = \{f^i \in \Omega \mid \delta(y^i,g_c) < \delta(y^i,g_{c'})\}$ (for c≠c')

$$\delta(\mathbf{y}^i,\mathbf{g}_c) = \sqrt{\sum_{j=1}^{J_i}\left(y^j - \sum_{l=1}^{H+M} B(\varsigma^c)\alpha_l^c\right)^2}$$



WDSA 2007

Cluster Characterization

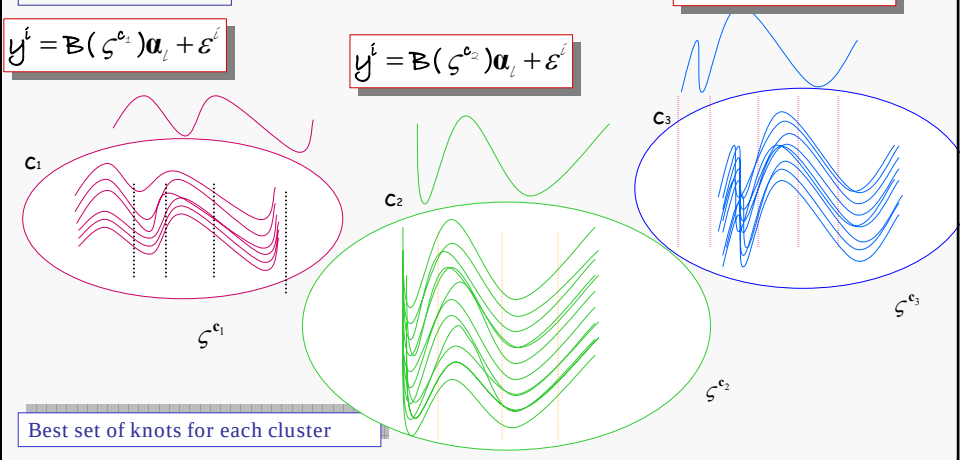
According the dynamical clustering procedure for each cluster we can characterize each cluster for sliding window:

Prototypal functions

$$y^i = B(\varsigma^{c_1})\alpha_l + \varepsilon^i$$

$$y^i = B(\varsigma^{c_2})\alpha_l + \varepsilon^i$$

$$y^i = B(\varsigma^{c_3})\alpha_l + \varepsilon^i$$



Best set of knots for each cluster

Cluster Characterization

Microprototypal functions are bound of variability obtained considering the maximum and minimum values of the functions in the cluster.

$$\underline{g}_c(t)(\zeta^c, \underline{\alpha}_l) = B(\zeta^c)\underline{\alpha}_l \quad c = 1, \ldots, C$$

$$\overline{g}_c(t)(\zeta^c, \overline{\alpha}_l) = B(\zeta^c)\overline{\alpha}_l \quad c = 1, \ldots, C$$

•It is able to monitor the variability of each cluster in different sliding window;

According to the Functional form characteristic of each cluster we can transform each curve inside the cluster according to the following function

$$\mathbf{x}^i = \mathbf{B}(\varsigma^c)\alpha_l$$

Cluster Characterization

Within variability

$$W_c(P_c, \hat{g}_c) = \sum_{c=1}^{C} \left\| x^i - \hat{g}_c(\zeta^c) \right\|^2$$

Prototypal function

$$f = \frac{1}{n} \sum_{c=1}^{C} n_c \hat{g}_c(\zeta^c)$$

Total variability

$$V = \sum_{c=1}^{C} W_c + B_c = \sum_{c=1}^{C} \sum_{i \in P_c} \left\| x^i - \hat{g}_c(\zeta^c) \right\|^2 + \sum_{c=1}^{C} n_c \| \hat{g}_c(\zeta^c) - f \|^2$$

Quality partition index

$$Q(P) = 1 - \frac{W(P, \hat{g}_c(\zeta^c))}{V}$$

Case study: Streaming curve data, propagation of sea waves

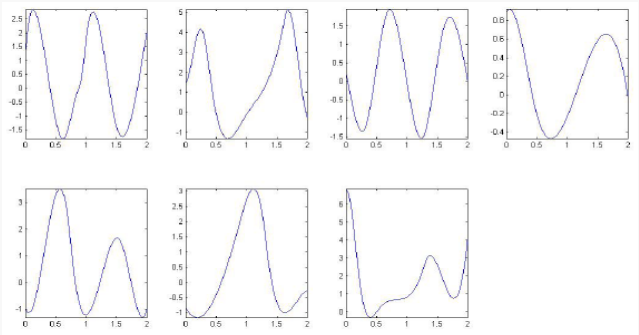
Since the detection of slope change point depends from the breaker type, several structure are considered. Expecially a nonlinear boundary-layer theory shows that strong streaming is possible for small viscosity.

As stayed before 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 novel future of our strategy is that it furnishes, classes of streaming curves with a set representative prototype of each cluster able to provide an helpful data summary .

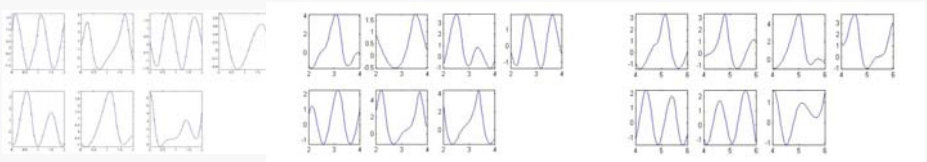
Main results

This is an outcome from a sliding window, the 7 prototype summarizes the change of the cluster structure, since the prototype differently from the other clustering algorithm represents the best model representing the clusters.



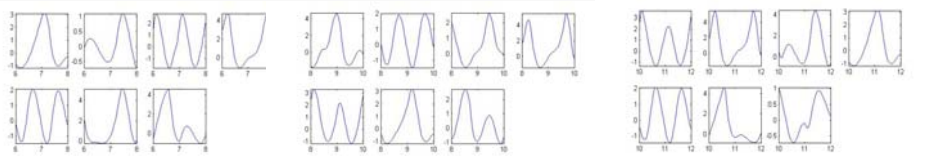
Main results

....for several sliding window .....
We have:



Best set of knots for each cluster:
a)(2.4619 2.7002 3.26 3.2611) **b)** (3.3999 3.4309) **c)**(2.6953 2.776 3.1741 3.2056) **d)** (2.3565 2.5427 3.2446 3.6031) **e)** (2.5485 3.2682 3.4521) **f)** (2.2949 2.3706 3.0578 3.6079 3.7615) **g)** (2.7067 3.3358 3.3962 3.5098)

Best set of knots for each cluster:
a)(4.7806 4.8514 5.266 5.3072) **b)**(4.6466 4.8937 5.1948) **c)** (4.7959 5.1706 5.2713 5.6391) **d)** (4.6969 5.2885 5.7828) **e)** (4.4585 4.7816 5.5609 5.6963) **f)** (4.6139 5.1081 5.5685) **g)**(4.4394 4.9362)



Best set of knots for each cluster:
a)(6.8749 7.3083 7.3497) **b)** (7.282 7.3457) **c)**(6.1601 6.6309 6.8592 6.9512 7.5039) **d)** (6.2921 6.3874 6.9486 7.6192 7.652) **e)** (6.2899 6.5799 7.2933 7.5288) **f)** (6.385 7.3463 7.3472) **g)** (6.5999 6.612 6.636 7.1053)

Conclusions

The proposed techniques seems capable to face the problem of extraction knowledge for curve datastream.

It exhibits superior performance in the perspective to catch meaningful characteristic of the cluster structure without considering the elements of the clusters but only the representative elements.

Perspectives

In the next step, it will be naturally interesting to compare our proposal with the existing techniques that have the main aim to extract information.

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).

WDSA 2007

Questions?

Elvira Romano *elvroman@unina.it*

Antonio Balzanella *balzanella2@alice.it*

Rosanna Verde *rosanna.verde@unina2.it*