

Monitoring spatially dependent data streams

Monitoraggio di flussi di dati spazialmente dipendenti

Antonio Balzanella and Rosanna Verde and Antonio Irpino

Abstract In many real world problems data streams mining deals with streams recorded by sensor networks located on some geographic area. Often, the placement of the sensors determines the presence of a spatial dependence among the data streams. This paper proposes a new method for monitoring the evolution of the spatial dependence among data streams. It is based on the variogram for histogram data which is used as synopsis for keeping a snapshot of the spatial dependence over time intervals. We propose an efficient way to compute, on-line, the variogram for histogram data and a measure for keeping track of evolutions in the spatial dependence.

Abstract *In numerosi ambiti applicativi, le tecniche di data stream mining sono utilizzate per l'analisi di flussi di dati acquisiti da reti di sensori georeferenziati. La collocazione spaziale dei sensori spesso determina la presenza di dipendenza spaziale. In quest'articolo, si propone un nuovo metodo per il monitoraggio del cambiamento della struttura di dipendenza spaziale tra flussi di dati basato sull'utilizzo del variogramma per dati ad istogramma, come sinopsi per tenere traccia della dipendenza spaziale nel tempo. Si propone un metodo efficiente per il suo aggiornamento on-line e una misura per valutare l'evoluzione della dipendenza spaziale nei dati.*

Key words: Data Stream Mining, CluStream, Variogram, Sensor data, histogram data

Antonio Balzanella
Second University of Naples, Viale Ellittico - Caserta - Italy, e-mail: antonio.balzanella@unina2.it

Rosanna Verde
Second University of Naples, Viale Ellittico - Caserta - Italy, e-mail: rosanna.verde@unina2.it

Antonio Irpino
Second University of Naples, Viale Ellittico - Caserta - Italy, e-mail: antonio.irpino@unina2.it

1 Introduction

Data Stream Mining (DSM) aims to extract knowledge from huge amounts of data which come from the repeated measurement of some variables. It assumes that the dataset is made by observations which become available over time at a very high frequency, evolving in an unpredictable fashion. Data with a time dimension should be processed with at least a partial temporal ordering, moreover the input data is not available for random access from disk or memory, but is rather represented as a continuous flow. A time window, that is a subsequence of a data stream, is generally available only once, thus, data has to be processed on fly at the speed in which it is recorded. Finally, algorithms that are applied for the analysis, need to change their behavior over time, according to the dynamic nature of data.

Often, data streams are recorded by sensor networks located on some geographic area. In many real world applications, such as climatology, this involves a dependence among data streams due to the spatial location of the sensors. In this sense, near located sensors tend to produce values which are similar while far sensors tend to produce more dissimilar values.

In dynamic frameworks where the monitored phenomenon evolves over time, the spatial dependence can it self evolve. The monitoring and discovering of evolutions in spatially dependence is the main topic of this paper.

To our knowledge, this task has not been dealt in the data stream mining literature. Methods for analyzing spatial data streams focus, mainly, on traditional data mining tasks (clustering, classification, summarization) on streams recording the position of moving objects [5][7]. Our challenge is, instead, the analysis of data produced by sensors having a fixed, known, spatial location. With this aim, we introduce a new tool for measuring the spatial dependence, called variogram for histogram data, and an efficient method for its on-line updating. We still introduce a measure for keeping track of evolutions in the spatial dependence.

2 Main notation and tools

Let $Y = \{Y_1, \dots, Y_i, \dots, Y_n\}$ be a set of n streams, with $Y_i = \{(y_i^1, t_1), \dots, (y_i^j, t_j), \dots\}$ be real valued ordered observations on a common discrete time grid $T = \{t_1, \dots, t_j, \dots\}$, with $t_j \subseteq \mathbb{R}$ and $t_j > t_{j-1}$. Each data stream Y_i is made by observations recorded by a sensor located at $s_i \in S$, with $S \subseteq \mathbb{R}^2$ be the geographic space.

The method we propose is based on splitting the incoming parallel streams into non overlapping windows, where each window, whose identifier is $w = 1, \dots, \infty$, is an ordered subset of T having size b . Each window, frames a subset $Y^w = \{Y_1^w, \dots, Y_i^w, \dots, Y_n^w\}$ where $Y_i^w = \{(y_i^j, t_j), \dots, (y_i^{j+b}, t_{j+b})\}$ is called subsequence.

A subsequence Y_i^w is summarized by a histogram $H_i^w = \{(I_{i,1}^w, \pi_{i,1}^w), \dots, (I_{i,L}^w, \pi_{i,L}^w)\}$ made by L weighted intervals (bins). Each histogram partitions the support $D_i^w = [y_i; \bar{y}_i]$ of a subsequence Y_i^w into a set of non overlapping intervals, or bins, so that:

$$D_i^w = \{I_{i,1}^w, \dots, I_{i,l}^w, \dots, I_{i,L}^w\}, \text{ where } I_{i,l}^w = [y_{i,l}, \bar{y}_{i,l}) \text{ (for } l = 1, \dots, L)$$

Since our method needs to compare histograms, the next section introduces the Wasserstein metric for histogram data.

2.1 Wasserstein metric for histogram data

The ℓ_p Wasserstein distance [9] is a distance function defined between the probability distributions of two random variables Y_1 and Y_2 , on a given metric space. Since Gini (1914) that introduced such measure in a discrete setting on the real line, several authors proposed this distance in different contexts of analysis. Therefore, we refer to this family of distances as Wasserstein distance. Denoting with F and G the distribution functions of Y_1 and Y_2 , and with F^{-1} and G^{-1} the corresponding quantile functions, Mallows [6] proposed a metric that is the ℓ_2 version of the Wasserstein distance which is expressed as follows:

$$d_W(Y_1, Y_2) := \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}$$

In particular, in our analysis we focus our attention on such ℓ_2 -norm distance because it can be interpreted as the Euclidean distance between quantile functions.

The main drawbacks are related to the inverse of the distribution functions which is impossible to do analytically for most distributions. As shown in [4][8], this problem can be addressed, when data are histograms, by introducing an exact and efficient way to compute this distance.

Given a histogram description H_i^w , with L be the number of weighted intervals (bins):

$$H_i^w = \{(I_{i,1}^w, \pi_{i,1}^w), \dots, (I_{i,l}^w, \pi_{i,l}^w), \dots, (I_{i,L}^w, \pi_{i,L}^w)\}$$

we define the quantities $\tau_{i,l}^w$ in order to represent the cumulative weights associated with the elementary intervals of H_i^w :

$$\tau_{i,l}^w = \begin{cases} 0 & l = 0 \\ \sum_{h=1, \dots, l} \pi_{i,h}^w & l = 1, \dots, L \end{cases} \quad . \quad (1)$$

Using (1), and assuming a uniform density for each $I_{i,l}^w$, the empirical distribution function is expressed as:

$$F_i^w(y) = \tau_i^w + \left(y - \underline{y}_{i,l}\right) \frac{\tau_{i,l}^w - \tau_{i,l-1}^w}{\bar{y}_{i,l} - \underline{y}_{i,l}} \quad \text{iff } \underline{y}_{i,l} \leq y \leq \bar{y}_{i,l}.$$

Then, the inverse distribution function is a piecewise function defined as follows:

$$F_i^{w-1}(t) = \underline{y}_{i,l} + \frac{t - \tau_{i,l-1}^w}{\tau_{i,l}^w - \tau_{i,l-1}^w} \left(\bar{y}_{i,l} - \underline{y}_{i,l}\right) \quad \tau_{i,l-1}^w \leq t < \tau_{i,l}^w.$$

To compute the distance between two histogram descriptions H_i^w and H_j^w we need to identify a set of common uniformly dense intervals. Let τ^w be the set of the cumulated weights of the two distributions $\tau^w = [\tau_0^w, \dots, \tau_l^w, \dots, \tau_q^w]$, where: $\tau_0^w = 0$, $\tau_q^w = 1$ and $\pi_l^w = \tau_l^w - \tau_{l-1}^w$. To solve the problem of finding a common set τ^w of cumulated weights associated with the quantiles of the two distributions, we consider equi-depth histograms. In this case, histograms H_i^w and H_j^w involved in the distance computation are characterized by the same set of weights $\pi_{i,l}^w = \pi_{j,l}^w = \frac{1}{L}$ and $q = L$.

The squared distance between the two histograms is computed as:

$$d_M^2(H_i^w, H_j^w) := \sum_{l=1}^q \int_{\tau_{l-1}^w}^{\tau_l^w} \left(F_i^{w-1}(t) - F_j^{w-1}(t)\right)^2 dt. \quad (2)$$

Each couple (τ_{l-1}^w, τ_l^w) permits to identify two uniformly dense intervals, one for i and one for j , having respectively the following bounds:

$$I_{i,l}^w = [\underline{y}_{i,l}, \bar{y}_{i,l}] \text{ and } I_{j,l}^w = [\underline{y}_{j,l}, \bar{y}_{j,l}].$$

Being $F_i^{w-1}(\tau_{l-1}^w) = \underline{y}_{i,l}$ and $F_i^{w-1}(\tau_l^w) = \bar{y}_{i,l}$ with $u = i, j$.

For each interval (bin) of the histogram, it is possible to compute the centers and radii, as follows:

$$c_{i,l} = (F_i^{w-1}(\tau_l^w) + F_i^{w-1}(\tau_{l-1}^w))/2 \quad r_{i,l} = (F_i^{w-1}(\tau_l^w) - F_i^{w-1}(\tau_{l-1}^w))/2.$$

Because intervals are uniformly distributed, we may express them in function of their centers and radii, and rewrite the equation (2) as follows:

$$d_W^2(H_i^w, H_j^w) = \sum_{l=1}^q \pi_l^w \left[(c_{i,l} - c_{j,l})^2 + \frac{1}{3} (r_{i,l} - r_{j,l})^2 \right]. \quad (3)$$

2.2 The variogram function

In the traditional geostatistics literature, a widely used tool for evaluating the spatial dependence is the variogram. Given a random process Y , it is defined as the variance of the difference between process values at two locations, across realizations of

the process [3]. If the process is stationary and isotropic, the variogram γ can be represented as a function of the distance $h = \|s_i - s_j\|$ between spatial locations. An unbiased estimator of the variogram, for a set of observations $y_i, i = 1, \dots, k$ located at s_i is the empirical variogram whose formal expression is the following:

$$\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{i,j \in N(h)} (y_i - y_j)^2 \quad (4)$$

where $N(h)$ is the set of observations such that $\|s_i - s_j\| = h$ and $|N(h)|$ is the number of pairs in the set.

In spatially dependent data, the value of the variogram function increases with the lag distance h until a limit is reached. This limit permits to identify the level beyond which the variability is no longer dependent on the spatial distance.

The variogram estimator cannot be computed at every lag distance h , due to the reduced availability of observations. This involves that the empirical variogram is not ensured to be valid so that in applied geostatistics, the empirical variograms are often approximated by model function ensuring validity [2].

The empirical variogram introduced above, can be still used as a pure exploratory tool for investigating spatial dependence in the data. In this sense, there is a distinction between the variogram model, estimated using the empirical variogram and than some fitting model function, and the experimental variogram which corresponds to the empirical variogram computed on the data without making formal assumptions on the process generating the data.

We are interested in this second use of the variogram function in order to monitor the evolution of the spatial dependence among the data streams.

Since the input data streams are represented as sequences of histograms, the next section will introduce the variogram for histogram data.

3 Experimental variogram for histogram data

Let $Y^w = \{Y_1^w, \dots, Y_i^w, \dots, Y_n^w\}$ be the set of $i = 1, \dots, n$ subsequences framed by the w -th time window and $s_i \in S$ (with $S \subseteq \mathbb{R}^2$) be the geographic locations of the data stream Y_i .

Each subsequence Y_i^w is represented by a histogram $H_i^w = \{(I_{i,l}^w, \pi_{i,l}^w), \dots, (I_{i,L}^w, \pi_{i,L}^w)\}$, made by L weighted intervals (bins).

We define the experimental variogram for histogram data consistently with the ℓ_2 Wasserstein metric introduced above and with its expression for histogram data in 3:

$$\gamma_H(h) = \frac{1}{|N(h)|} \sum_{i,j \in N(h)} (d_W^2(H_i^w, H_j^w)) \quad (5)$$

where $N(h)$ is the set of observations such that $\|s_i - s_j\| = h$ and $|N(h)|$ is the number of pairs in the set.

The variogram $\gamma_H(h)$ is nonnegative since it is the average of squared distances and shares the characteristics of the experimental variogram for traditional scalar data.

For irregularly spaced data where there are not enough observations exactly separated by h , $N(h)$ is modified to $(s_i, s_j) : \|s_i - s_j\| \in (h - \varepsilon, h + \varepsilon)$, with $\varepsilon > 0$. This involves that $\gamma_H(0) \geq 0$ (nugget effect).

According to the provided definition of variogram for histogram data, we can measure the spatial dependence between distributions (histograms) into a time window. If there is spatial dependence, we expect that near sensors tend to have lower values of average distances while far sensors tend to be more different so that $\gamma_H(h)$ is an increasing function.

Our aim is to monitor the evolution of the variogram for histogram functions over time windows. With this aim, $\gamma_H(h)$ has to be updated with the arrival of new windows of data.

In the next section we introduce an effective way for updating the variogram for histogram functions reducing the computational effort.

4 On-line updating of the variogram for histogram functions

Keeping track of the evolutions in spatial dependence requires the monitoring of changes in the variogram for histogram functions. Our strategy is to update the variogram every time a new batch of subsequences becomes available and store a snapshot of the variogram at predefined time stamps using a tilted time frame approach. The detection of change points is performed by comparing the variograms coming from the snapshots.

The variogram construction is based on computing the average of pairwise distances at each lag distance h . The set of pairs involved in the average computation depends only on the spatial location of the sensors which record the data. Such locations are fixed and known apriori so that at each window always the same pairs participate to the computation of the variogram value.

The variogram function $\gamma_H^w(h)$ for the window w , starting from the variogram $\gamma_H^{w-1}(h)$ at the window $w - 1$ can be computed as follows:

$$\gamma_H^w(h) = \frac{1}{2|N^w(h)|} \left(\gamma_H^{w-1}(h) |N(h)| + \sum_{i,j \in N(h)} (d_W^2(H_i^w, H_j^w)) \right) \quad (6)$$

where $N^w(h) = N^{w-1}(h)$ is the set of observations such that $\|s_i - s_j\| = h$ and $|N^w(h)|$ is the number of pairs in the set.

The previous expression indicates how to update the variogram at each lag h by computing new average distances.

This methods for updating the variogram at each window still requires the computation of the distance between every pair of subsequences at a spatial distance h . To make more efficient this step, we use the micro-clustering method for histogram data, proposed in [1], on each data stream Y_i .

The cited method performs the analysis of a data stream in order to obtain a set of low variability micro-clusters which are summarized by histogram centroids. Consistently with the approach proposed in this paper, the micro-clustering method splits the incoming data into windows and allocates the histograms to the clusters using the adapted Wasserstein distance to work with histogram data.

We propose to replace the computation of the pairwise distances in the variogram with the distance between micro-cluster centroids. The algorithm can be summarized as follows:

- For Each window w
 - For Each subsequence Y_i^w
 - Compute the histogram H_i^w
 - Allocate H_i^w to a micro-cluster
 - End
 - VARIOGRAM UPDATING
 - For Each lag distance h
 - Detect all the pairs H_i^w, H_j^w such that $\|s_i - s_j\| = h$
 - For Each pair H_i^w, H_j^w
 - detect the micro-cluster centroids to which H_i^w and H_j^w have been allocated
 - End
 - Update the variogram value γ_H^w computing the distance between the centroids to which H_i^w and H_j^w have been allocated
 - End
- End

The distances between micro-cluster centroids can be stored into a lookup table so that it is not needed to compute them every time. This makes the variogram updating computationally faster.

In order to evaluate the evolutions in spatial dependence, the variogram for histogram data can be stored at predefined time stamps and computed again on the incoming batches. The comparison can be performed graphically or computing the Euclidean distance between two variogram functions as follows:

$$EVO = \sqrt{\sum_h (\gamma_H^w(h) - \gamma_H^{w-k}(h))^2} \quad (7)$$

The computation of the measure EVO is feasible since the data streams have a fixed spatial location so that the compared experimental variograms are computed for the same values of the lag h

5 Experimental results

We have made some preliminary test for evaluating the performance of the proposed strategy in keeping track of the spatial dependence among sensor data using a public dataset available at <http://db.csail.mit.edu/labdata/labdata.html>.

The dataset collects the records of 54 sensors placed at the Intel Berkeley Research lab between February 28th and April 5th, 2004. Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. Data was collected using the TinyDB in-network query processing system, built on the TinyOS platform. The dataset includes the x and y coordinates of sensors (in meters relative to the upper right corner of the lab).

We have analyzed the temperature records of each sensor so that we have a set of 54 time series each on made by 65000 observations.

In order to run the test, we have set the size of each window to $s = 200$ and the number of bins for each histogram to $L = 10$. In fig.1 and in fig.2 we have plotted the variogram for histogram data obtained by processing, respectively, the first 32500 time stamps and the latest 32500 time stamps. Looking at the plots, we can note a change in the spatial dependence between the first batch of data and the second one, however both highlight the presence of a spatial dependence since the two plots tend to increase with the lag distance h .

6 Conclusions

In this paper we have introduced a strategy for monitoring the spatial dependence of data recorded by spatially located sensors. To reach this aim, we have, at first,

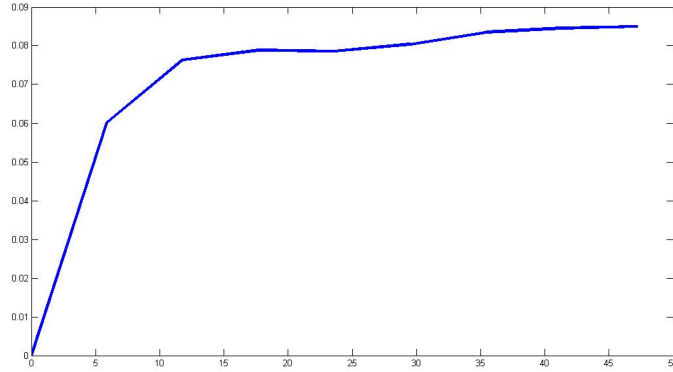


Fig. 1 Variogram function for histogram data computed on the first 32500 time stamps

introduced the variogram for histogram data and, then, a way to update it with the flowing of data. Preliminary results confirm the effectiveness of the method, however, in future works, we will explore the possibility to reduce the computational effort required for keeping an updated variogram over time.

References

1. Balzanella A., Rivoli L., Verde R. Data stream summarization by histograms clustering. In: Statistical Models for Data Analysis. Eds: Giudici; Ingrassia; Vichi, Springer. ISBN:978-3-319-00031-2 (2013)
2. Chiles, J. P., Delfiner P.: Geostatistics, Modelling Spatial Uncertainty, Wiley-Interscience (1999)
3. Cressie, N.: Statistics for spatial data, Wiley Interscience (1993)
4. Irpino, A., Verde, R., Lechevallier, Y.: Dynamic clustering of histograms using Wasserstein metric. In: COMPSTAT, pag.869876 (2006)
5. Ling-Yin Wei, Wen-Chih Peng: An incremental algorithm for clustering spatial data streams: exploring temporal locality. In: Knowledge and Information Systems Volume 37, Issue 2, pp 453-483 (2013)
6. Mallows, C. L.: A Note on Asymptotic Joint Normality. In: The Annals of Mathematical Statistics, **43** (2), 508515. (1972)
7. Qiang Ding, Qin Ding, William Perrizo: Decision tree classification of spatial data streams using Peano Count Trees. In: Proceedings of the 2002 ACM symposium on Applied computing, March 11-14, Madrid, Spain (2002)
8. Verde, R., Irpino, A.: Dynamic Clustering of Histogram Data: Using the Right Metric. In: P. Brito, G. Cucumel, P. Bertrand, F. Carvalho (Eds.), Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, chap. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, 123-134 (2007)
9. Wasserstein, L. Markov processes over denumerable products of spaces describing large systems of automata. In: Prob. Inf. Transmission, **5**, 47-52 (1969)

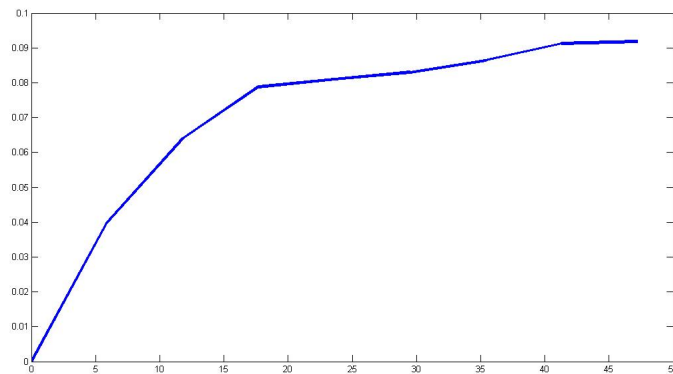


Fig. 2 Variogram function for histogram data computed on the latest 32500 time stamps