Data stream summarization by on-line histograms clustering

Antonio Balzanella[*], Lidia Rivoli[**]
Rosanna Verde [*]

[*]Second University of Naples - Via del Setificio 15 - San Leucio - 81100 Caserta - Italy
antonio.balzanella@gmail.com,
[**]University of Naples Federico II - Complesso universitario di Monte S. Angelo Napoli
lidia.rivoli@unina.it

In recent years, a wide number of applicative fields is generating continuous, potentially unbounded data streams. The analysis of such kind of data is constrained by the impossibility to store the whole dataset and by the need to provide the results as soon as possible in order to support the decisions.

When we are dealing with highly evolving data, an important challenge is to discover summaries able to highlight the main concepts which characterize the analyzed phenomenon.

In this context, we introduce an efficient strategy which provides, as output, a set of histograms to summarize the main concepts emerging in an evolving data stream.

A datastream $Y = \{(y_1, t_1), (y_2, t_2), \ldots, (y_\infty, t_\infty)\}$ is a set of real valued ordered observations on a discrete time grid $T = \{t_1, ..., t_2, ...t_\infty\} \in \Re$. From Y, it is possible to get a data batch $Q_i = \{(y_l, t_l), \ldots, (y_j, t_j), \ldots, (y_n, t_n)\}$ with $i \in \Im$, where $\Im$ is the unbounded set of all the ordered subsets of Y such that $Q_i \bigcap Q_{i+1} = \emptyset$. The size of Q_i is $N = n - l$.

We can synthesize the data by a histogram as follows. Let $S = [\underline{y}; \overline{y}]$ be the support of a data batch Q_i. The observations in Q_i are partitioned into a set of contiguous intervals (bins) $\{I_{1i}, \ldots, I_{ki}, \ldots, I_{Ki}\}$ where $I_{ki} = [\underline{y}_{ki}; \overline{y}_{ki})$ and $\bigcup_{k=1}^{K} I_k = [\underline{y}; \overline{y}]$. To each interval I_{ki} we associate the relative frequency f_{ki}, which is the number of elements of Q_i in $[\underline{y}_{ki}; \overline{y}_{ki})$ normalized to N.

Histogram construction requires the definition of the size and number of intervals. In this paper we make reference to equi-depth histograms where the range of observed values is divided into K intervals such that each interval include the same numbers elements.

The aim of this paper is to detect a set of summaries $G = \{g_1, \ldots, g_z, \ldots, g_Z\}$ which represents the histograms H_i associated to the batches of data Q_i. The strategy we introduce to reach this aim, is made by an on-line step and on an off-line step. The former, allows to get a set of synopsis of the stream, the latter, starts from the results of the on-line step to produce the final set of summaries G.

It is a variation of the CluStream algorithm in (Aggarwal et al., 2003). In particular, the on-line step looks for synthesis of non overlapping batches of data by means of a set of size $C >> Z$ of specific structures named micro-clusters. A micro-cluster stores a prototype g_c, the number of allocated histograms n_c.

Every time a new batch of data Q'_i is available and the associated histogram H'_i is constructed, the distance between H'_i and the prototype $g_c, \forall c = 1, \ldots, C$ of each micro-cluster is computed. If the distance to the nearest prototype is lower than a fixed threshold value, th,

H_i' is allocated to it and the statistics of the micro-cluster are updated. If no prototype is at a distance lower than th, a new micro-cluster is generated having H_i' as prototype and $n_c = 1$.

In order to compare the histograms and to compute the prototype of each micro-cluster, we need to introduce a suitable distance function.

We propose to use the Wasserstein distance as shown in Verde and Irpino (2007).

Let F and G be two distributions and F^{-1} and G^{-1} their the quantile functions. It is possible to define the Mallow's (Mallow (1972)) distance in L^2 derived from the Wasserstein metric as follows:

$$d_M(F,G) := \sqrt{\int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt} \quad (1)$$

According to Verde and Irpino (2007), since each interval of the histogram may be expressed as a function of the centers and of the radii $c + r(2t-1)$ for $0 \leq t \leq 1$, the (1) can be calculated much easier.

Furthermore, the use of Wasserstein metric allows to find the prototype of each micro-cluster as a histogram that is barycentric with respect to the elements of the cluster. This is obtained as the average of the centers and of the radii of each interval of the histograms in the cluster.

Finally, from the on-line updated micro-clusters, it is possible to discover the final set of summaries G through a clustering procedure on the micro-clusters that is a variation of the *k-means* algorithm. The output will be the final set of summaries $G = \{g_1, \ldots, g_z, \ldots, g_Z\}$.

References

Aggarwal, C. C., J. Han, J. Wang, and P. S. Yu (2003). A framework for clustering evolving data streams. In *VLDB '2003: Proceedings of the 29th international conference on Very large data bases*, pp. 81–92. VLDB Endowment.

Verde, R. and A. Irpino (2007). Dynamic clustering of histogram data: using the right metric. In *in: Selected Contributions in Data Analysis and Classification*, pp. 123–134. Springer.

Résumé

In recent years a wide range of applications generates potentially unbounded data streams. When we are dealing with highly evolving data, summaries able to highlight the main concepts in the monitored phenomenon are needed. In this paper we introduce a new strategy able to summarize the data flow through a set of histograms. It is a clustering procedure where the prototypes of the clusters are properly detected histograms.