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,