Mining Behavior Patterns in Streaming Multivariate Data

Klaus Mueller, Computer Science Department, Stony Brook University
Wei Xu, Computational Science Initiative, Brookhaven National Lab

Abstract: We present ideas and initial work on visual behavior mining in streaming multivariate data. A *behavior* in this context is a finite time-varying pattern of a single variable that can be stored into a library and can be compared with patterns of other variables to determine dependencies, correlations, and possibly causations. We have recently constructed a prototype, called *StreamVis*ND that implements some of our ideas. Our prototype couples analysis with interactive visualization and enables domain experts to apply their domain knowledge and intuition to mine, hypothesize, confirm, and reject multivariate behavior relationships in an interactive manner. We applied our visual analytics system in an environmental pollution diagnostics setting as a test case and have obtained encouraging results.

1. Introduction

Streaming data present a set of unique challenges because of the constraints associated with the large volume of continuously arriving data. Some of these constraints are (1) concept drift – the evolution of data over time, (2) one pass constraint – the massive data only allow processing when they stream by, and (3) massive-domain constraint – the data is so large that we cannot possible store everything. Due to these constraints virtually all streaming methods embed some kind of online synopsis construction approach into the mining process. This online synopsis is then used later in the miming process. A well-known example is reservoir sampling which guaranties certain probabilities for any stream point to be included into the synopsis structure.

We have been interested to mine *time patterns* in *multivariate* streaming data, also called *behaviors*. The particular aim is to detect and visualize when variables undergo concept drift and change their behavior relationships to other variables. This can uncover causal relationships or other dependencies. Storing behaviors in place of individual data can also help to overcome the massive data constraint. One can simply store an index into a reservoir of behaviors.

Apart from the underlying analytics we are also especially interested in effective visualization methods that can communicate these changing relationships in the context of the variables themselves. This requires new multivariate visualization methods and human-computer interaction paradigms that can deal with time-varying patterns which furthermore change over time.

2. Our StreamVisND Prototype

We have begun addressing these challenges in a framework we call *StreamVs*ND. An annotated screenshot of its interface is shown in Fig 1. StreamVsND runs on any modern web browser and is therefore platform independent. The *Temporal Attribute Display* shows the relation of the attributes in terms of their behavior measured within a certain time interval over time. Essentially this display can visualize concept drift in the behavior of the individual variables. Each colored line is due to one attribute and when two attribute lines are close this means that the behavior within the associated time slice is similar. We construct the plot by first subdividing the time series into equal-sized time intervals (the time slices) of length T, where T is the number of discrete time steps of the slice. Then, for each slice S_t and attribute A_j we construct a slice-attribute vector SA_{it} of length T which holds the T values A_i has within the given S_t, ordered by time stamp. Then we use pre-seeded MDS (Multidimensional Scaling) to arrange the variables along each vertical line and connect the points of each variable with a spline.

The *Time Slice Similarity Plot* visualizes the similarity of the time slices in terms of the attribute values – the darker time slices are those with higher time stamps. The *Dynamic Local Change* plot visualizes the local (transient) changes via MDS (Multidimensional Scaling), now by ways of a dynamic layout where

the local changes of the points are visualized with streak lines. The *Streamgraph and Time Slice Selector* enables users to (1) select a time window and so restrict the number of points shown, (2) click on a point and see the variable's proportions as a pie chart directly inserted into the Time Slice Similarity Plot, and (3) draw a rectangle and visualize the corresponding slices in the Temporal Attribute Relation Display.

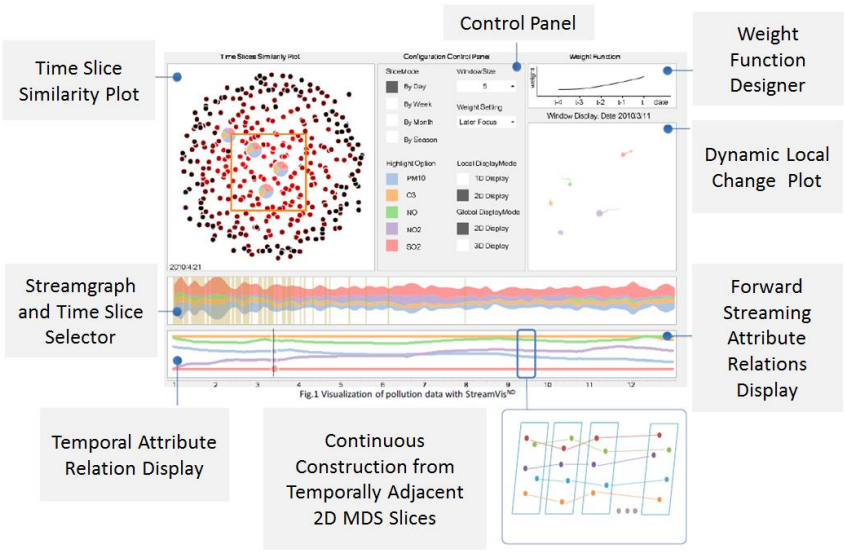

Fig. 1: Interface and components of our StreamVis[ND] prototype.

3. Case Study

In [1] we reported on the application of StreamVis[ND] to urban pollution data acquired by some of our collaborators in 2013. They measured concentrations of PM10, O3, NO, NO2 and SO2 over time at various locations in the city. The environment experts we worked with were especially interested in determining which two days had similar pollution profiles, which days were outliers, what were the relations of the pollutants over time, etc. In the Temporal Attribute Display we can easily observe that PM10 and NO2 have a rather similar time-behaviour from the middle of March to the middle of June. We can also find that SO2 does not have a close temporal correlation with NO and O3. On the other hand, in the Time Slice Similarity plot, by ways of the node coloring, we observe that early days (bright) are similar and aggregate in the display center, while later days (dark) are more dissimilar and map in a distributed manner into the periphery.

4. Possible Avenues for Future Work

There are several avenues for future work which will make the system more robust and more general applicable to other streaming data applications. Some of these are listed in the following:

Attribute Time Relation Display: So far we focused on visualizing temporally changing relationships among variables. But this is only part of the story. Often it is also interesting at what time periods one or more variables exhibited similar behaviors. This can be visualized by exchanging time with attributes in the Temporal Attribute Relation Display and using a similar similarity-based layout optimization scheme.

More distance functions: The measure of similarity is highly dependent on the distance function used for assessment. So far we used the traditional Euclidian distance, but other metrics such as correlation, our own structural similarity [2], dynamic time warping, auto regression, etc. might show features better.

Computing the best length of the time slice: Finding the inherent periodicity of the data – possibly at multiscale – to determine the lengths of the time slices is crucial for good matching. Frequency and wavelet analysis are promising instruments here and we plan to study these and other methods.

References

[1] S. Cheng, Y. Wang, D. Zhang, Z. Jiang, K. Mueller, "StreamVisND: Visualizing Relationships in Streaming Multivariate Data," *IEEE Visualization,* Chicago, IL, October, 2015 (won Honorary Mention Award).

[2] J. Lee, K. McDonnell, A. Zelenyuk, D. Imre, K. Mueller, "A Structure-Based Distance Metric for High-Dimensional Space Exploration with Multi-Dimensional Scaling," *IEEE Trans. on Visualization and Computer Graphics,* 20(3): 351-364, 2014.