GraphScan: Detection and Analysis of Free-Form Spatio-Temporal Patterns

Lukasz P. Wawrzyniak (lwawrzyn@uoguelph.ca)

Department of Computing and Information Science, University of Guelph, Ontario, Canada

OBJECTIVE

This paper proposes an efficient and flexible algorithm applicable to spatio-temporal aberration detection in public health data.

BACKGROUND

By capturing the spatio-temporal organization of the data using a graph, GraphScan avoids the challenges associated with trying to "fit" incoming data into moving windows of predefined shapes and sizes. Whereas the popular space-time permutation scan statistic [1] attempts to find clusters within space-time volumes of predefined shape, GraphScan employs no such preconceptions about the form of the clusters. Instead, clusters are allowed to "evolve" freely to better reflect the structural properties of the data. Moreover, GraphScan is capable of tracking possible causal relationships between spatio-temporal events.

METHODS

The spatial organization of geographical regions is represented using a planar graph. Each region is a vertex and two vertices are connected by an edge if and only if the two corresponding regions share a common boundary. In order to incorporate the temporal dimension into the analysis, the spatial graph is extended into a space-time graph. Just as every geographical region has a spatial neighbourhood (composed of all the regions adjacent to it), it also has a temporal neighbourhood. The temporal neighbourhood of a region at time t is composed of its "previous self" (i.e., the region at $t - 1$) and it's "future self" (i.e., the region at $t + 1$). One may think of the space-time graph as a structure composed of multiple layered and interconnected instances of the spatial graph. Thus, the space-time graph encapsulates both the spatial organization of regions as well as the time series associated with each region.

Vertex weights are assigned according to the level of "activity" (e.g., count of GI-related ED visits) for the corresponding region and time period. Clustering is the task of finding sets of connected vertices that simultaneously exhibit unusually high levels of activity. It can be formulated in terms of identifying the connected components of the space-time graph.

Now, consider a dataset consisting of 80,000 records from 150 regions collected over a period of five years (this is similar to several datasets presently at the author's disposal). The space-time graph is immense. Astronomical storage requirements aside, the running time of any clustering algorithm on a graph of this size would easily transcend the realm of practicality. To overcome this problem, an incremental clustering algorithm was developed that handles data on a day-by-day basis. In essence, the program consists of two major components: The *scanner*, which detects spatial clusters in the most recent data (the "current" date), and the *event tracker*, which maintains information about previously seen spatial clusters and groups overlapping spatial clusters into spatio-temporal clusters called *events*.

It is important to note that the existence of a cluster is not necessarily related to a lapse in public health. Clusters occur naturally in the data and may be caused by many things, including pure chance. One factor that might attach some significance to a spatial cluster is its persistence over time. If the cluster was indeed formed by chance, it is likely to dissipate rather quickly. If it persists, it may warrant further investigation into its true causes. Useful quantitative properties of spatio-temporal clusters include duration, coverage, total size, density, and stationarity (or drift, over time). These properties can be used to distinguish between mundane and suspicious activity patterns.

RESULTS

GraphScan was employed in retroactive analysis of (recent) historical ED data. The results seem promising: it was found that endemic activity patterns result in small, fleeting clusters. It can be conjectured that an outbreak of any significant magnitude would be readily discernible from the background signal.

Presently, a rigorous testing framework is being developed to assess the viability of GraphScan as an aberration detection technique. The generation of realistic synthetic (baseline and outbreak) data, identification of key quantitative properties of spatio-temporal events, and parameter space exploration are being vigorously addressed. Results of these efforts will be available shortly, and certainly before the conference commences.

CONCLUSIONS

GraphScan is more than a tool for detecting spatio-temporal clusters. An exciting avenue of this research is the analysis of *event chains*, i.e., the perceived links between spatio-temporal events.

REFERENCES

[1] Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F., A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine, 2:216-224, 2005.

Further Information:
Lukasz Wawrzyniak, lwawrzyn@uoguelph.ca