Penalized fast subset scanning

Skyler Speakman¹, Sriram Somanchi¹, Edward McFowland III¹, Daniel B. Neill¹,*

1. Event and Pattern Detection Laboratory, Carnegie Mellon University
*Contact author: neill@cs.cmu.edu

Keywords: Likelihood ratio statistic, Scan statistic, Pattern detection, Disease surveillance

We present the Penalized Fast Subset Scan (PFSS), a new and general framework for scalable and accurate pattern detection. PFSS enables exact and efficient identification of the most anomalous subsets of the data, as measured by a likelihood ratio scan statistic. However, PFSS also allows incorporation of prior information about each data element’s probability of inclusion, which was not previously possible within the subset scan framework. PFSS builds on two main results: first, we prove that a large class of likelihood ratio statistics satisfy a property which allows additional, element-specific penalty terms to be included while maintaining efficient computation. Second, we prove that the penalized statistic can be maximized exactly by evaluating only $O(N)$ subsets. As a concrete example of the PFSS framework, we incorporate “soft” constraints on spatial proximity into the spatial event detection task, enabling more accurate detection of irregularly-shaped spatial clusters of varying sparsity. To do so, we develop a distance-based penalty function which rewards spatial compactness and penalizes spatially dispersed clusters. This approach was evaluated on the task of detecting simulated anthrax bio-attacks, using real-world Emergency Department data from a major U.S. city. PFSS demonstrated increased detection power and spatial accuracy as compared to competing methods while maintaining efficient computation. This work was partially supported by National Science Foundation grants IIS-0916345, IIS-0911032, IIS-0953330, and GRFP-0946825. The full paper is available as Speakman et al. [1].

References