III : Small : Fast Subset Scan for Anomalous Pattern Detection

This work will develop new methods for fast and scalable detection of *anomalous patterns* (subsets of the data that are interesting or unexpected) in massive, multivariate datasets. We focus on real-world applications where we must detect complex, subtle, and probabilistic patterns that are difficult to spot with existing techniques, such as an emerging disease outbreak or a pattern of smuggling activity.

Our proposed work is based on two key insights. First, the pattern detection problem can be framed as a *search* over all subsets of the data, in which we define a measure of the "anomalousness" of a subset and maximize this measure over all potentially relevant subsets. We have incorporated this insight into a general "subset scan" framework for pattern detection. Second, and more surprisingly, we have discovered that, for many spatial detection methods (including Kulldorff's spatial scan statistic and many recently proposed variants), we can perform an exact search which efficiently maximizes our measure of anomalousness over all subsets of the data. We propose to explore this new combinatorial optimization method, investigate how it can be extended to constrained subset scans and to more general multivariate pattern detection problems, and examine how it can be incorporated into our subset scan framework, enabling us to create a variety of fast, scalable, and useful methods for anomalous pattern detection.

Intellectual Merit. We will develop, implement, and evaluate a general probabilistic framework for efficient detection of anomalous patterns in both spatial and non-spatial datasets. The proposed work will address these challenging and important research questions:

- How can we define a useful measure of the "anomalousness" of a subset of the data, and efficiently optimize this measure over all subsets to find the most anomalous patterns?
- What are the necessary and sufficient conditions for a set function $F(S)$ to satisfy the "linear-time subset scanning" (LTSS) property, enabling exact unconstrained optimization of $F(S)$ over all 2^N subsets of N records while only requiring $O(N)$ subsets to be evaluated?
- How can we extend our fast subset scanning methods to general multivariate datasets, and incorporate search constraints such as proximity, connectivity, and self-similarity?
- How can we deal with uncertainty about the effects of an anomalous pattern by efficiently searching over subsets of "input" and "output" attributes as well as subsets of records?

The ability to efficiently find the most anomalous subsets of a massive dataset, with or without additional constraints, will provide a *qualitatively* new approach for scalable pattern detection.

Broader Impact. Development and testing will be prioritized in three areas: early detection of disease outbreaks, detecting illicit container shipments, and identifying anomalous trends in social networks. These applications will demonstrate the value of our methods across a wide spectrum of domains. Through the PI's existing collaborations, the algorithms will be incorporated into deployed systems for health and crime surveillance that contribute directly to the public good.

The PI has been developing novel and computationally efficient machine learning methods to solve real-world pattern detection problems, and publishing in top conferences and journals, since 2002. His lab has over 5 years of history offering free machine learning software, and the software implementations of all algorithms developed through this grant will be made publicly available.

The bulk of the requested funding will go to training graduate students who will become the next generation of researchers to explore new methods for anomalous pattern detection. This effort will build collaborations among students and researchers in the emerging interdisciplinary field of Machine Learning and Public Policy (MLPP) as part of a new initiative spearheaded by the PI, including a new joint PhD program in MLPP, and development of new MLPP courses and seminars.

Key Words: anomalous patterns; pattern detection; fast subset scan; scan statistics; optimization.