III : Large : Discovering Complex Anomalous Patterns

Many of the most interesting and valuable discoveries from data arise not from the evaluation of single records, but from identifying a set of records that are anomalous in some interesting way. Together these records may indicate for example the emergence of a disease outbreak. In scientific domains, they may represent new phenomena waiting to be discovered.

We view pattern discovery as an interactive process between a discovery system and a human user who has expertise in the domain. We propose to develop an integrated framework of probabilistic methods that allow the system to work interactively with a user in detecting, characterizing, explaining, and learning anomalous patterns over groups of records. Our focus is on the many situations where the data (and the probabilistic patterns we seek to discover) are not appropriate for using other existing techniques, such as graph mining or frequent sets.

Our proposed methods will search over arbitrary subsets of records and evaluate how well they are predicted by known probabilistic models, which represent expected stochastic patterns of the data that are potentially quite complex. They will detect both *known patterns*, subsets corresponding to some known and modeled pattern type of high relevance to the user, and *unknown patterns*, subsets that are highly unexpected given the background data or any known pattern type. The system will assist the user in understanding and modeling the discovered, currently unknown anomalies, so that each will be identified as a known pattern when encountered in the future.

Intellectual Merit. We will develop, implement, and evaluate a general, comprehensive, and widely applicable probabilistic framework for pattern discovery. The proposed work will address these challenging and important research questions:

- How can machine learning concepts such as classification and anomaly detection be generalized to consider groups of records rather than single records?
- How can a detection algorithm simultaneously detect and differentiate between known and currently unknown pattern types?
- How can an algorithm explain clearly to a user what pattern was found and why?
- How can an algorithm learn new pattern types through feedback from a user?

The ability to detect, characterize, explain, and learn patterns from groups of records in massive datasets will provide a *qualitatively* new approach for advancing discovery of knowledge from data.

Broader Impact. Although the applications for these algorithms are innumerable, development and testing will be prioritized in the areas of patient care in the intensive care unit (ICU) and aircraft fleet maintenance. Through the project investigators' existing collaborations, the algorithms will also be used during the project in other areas including food safety, scientific discovery in astronomy sky surveys, and detection of geographic hot-spots of criminal activity. Together, these applications will demonstrate the methods' value across a wide spectrum of domains and tasks.

The PIs have been working closely together since 2001, developing novel and computationally efficient machine learning algorithms, and publishing in top conferences and journals. Their 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 at www.autonlab.org.

The bulk of the requested funding will go to training Ph.D. students who will become the next generation of researchers to develop new algorithms for discovering complex phenomena in data. This effort will build interdisciplinary collaborations among researchers in computer science, public policy, health care, and biomedical informatics. This will occur through the direct training of Ph.D. and postdoctoral trainees, the inclusion of new insights in graduate courses and seminars taught by the investigators, and as part of a new joint Machine Learning/Public Policy Ph.D. program.

Keywords: anomalous patterns; pattern discovery; probabilistic models; incremental learning.