

Symposium on Machine Learning for Anomaly Detection

Final Report for NSF Grant IIS-0442128

Stephen Bay and Pat Langley
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

The Symposium on Machine Learning for Anomaly Detection was held at Stanford University on May 22 and 23rd, 2004. Over 30 researchers attended the meeting, of which 11 participants, all known for their work in this area, presented talks on their recent results. The symposium fostered discussion between scientists working on anomaly detection in disparate domains and provided a forum where they could meet to share their experiences and approaches. The role of machine learning was the central theme of the meeting and the speakers discussed many common issues such as representation, cost functions, and controlling false positives.

We organized the symposium into talks and discussions over two consecutive days. We also dedicated a session of the second day to the issues of improving research on anomaly detection within the broader machine learning community. Below we summarize the contents for each speaker's presentation in the order given. This information, together with slides and references to what each author judged to be his or her most relevant paper, can be found on the symposium Web site at <http://c11.stanford.edu/symposia/anomaly/>.

1. ACTIVITY MONITORING: ANOMALY DETECTION AS ON-LINE CLASSIFICATION. *Tom Fawcett* (HP Labs) argued that many anomaly detection problems can be viewed as involving on-line stream classification. He presented a basic analysis framework that covered problems in diverse domains such as fraud detection, computer intrusion detection, network performance monitoring, epidemic detection, and news story tracking. A central issue in the framework is the development of evaluation metrics that account for the temporal nature of the problem.
2. LEARNING MODELS FOR DETECTING ANOMALIES IN TIME SERIES. *Philip Chan* (Florida Institute of Technology) reported a method for detecting anomalies in NASA shuttle data. His approach is based on segmenting the time series into states, using classification rules to create descriptions of the states, and modeling transitions between states with a finite-state automaton.
3. A FRAMEWORK FOR DETECTING ANOMALOUS REGIMES IN TIME-SERIES DATA. *Stephen Bay* (Stanford University) presented an approach to detecting anomalous regimes, which he defined as time periods with unusual causal relationships among the observed variables. He demonstrated how such a tool could be used in monitoring complex systems and in exploratory investigations to discover new phenomena.
4. TOWARDS PARAMETER FREE ANOMALY DETECTION. *Eamonn Keogh* (UC Riverside) presented a nearly parameter-free algorithm for detecting anomalies in time-series data. The central idea is to estimate the increase in Kolmogorov complexity with standard data-compression algorithms when comparing two segments of data. Large increases suggest that the series are anomalous with respect to each other.

5. DETECTING UNIQUE DOCUMENTS VIA VISUALIZATION. *Kazumi Saito* (NTT) presented an algorithm for finding unusual documents with visualization. His method is based on embedding the documents into a low-dimensional vector space that maximally preserves probabilistic relationships. He showed how, once each document has been transformed into this new space, the entire set of documents can be easily viewed and manipulated to detect anomalies.
6. DETECTING BIO-TERRORIST ATTACKS BY MONITORING MULTIPLE STREAMS OF DATA. *Galit Shmueli* (University of Maryland, College Park) focused on methods for early detection of bio-terrorist attacks or other outbreaks from non-traditional data sources such as over-the-counter medication sales, nurse hotlines, or even Web searches. Her approach involved using wavelet transforms to decompose the data into multiple time scales.
7. WHAT'S STRANGE ABOUT RECENT EVENTS. Weng-Keen Wong (Carnegie Mellon University) presented WSARE, an algorithm for early detection of disease outbreaks that monitors a variety of spatial, temporal, demographic, and symptomatic information. One of WSARE's unique features is its use of a Bayesian network to produce a baseline trend that adaptively accounts for temporal variations such as seasonal effects and day of the week variations.
8. SPATIAL OUTLIER DETECTION AND APPLICATIONS. Chang-Tien Lu (Virginia Tech) discussed the problem of detecting spatial outliers in geographical data and presented several approaches for single and multiple attribute data. He also described ways to find and track unusual spatial regions which are based on wavelet decompositions that analyze the data at multiple scales.
9. ANOMALY DETECTION APPLICATIONS IN EARTH SCIENCE. Mark Schwabacher (NASA Ames Research Center) reviewed the data generated by NASA's Earth Observing Satellites and discussed why earth scientists are interested in analyzing them to discover anomalies. He presented an algorithm for finding point anomalies that scales well to the massive data volumes produced, and then discussed his initial work on finding region-based anomalies.
10. A DECISION-THEORETIC, SEMI-SUPERVISED MODEL FOR INTRUSION DETECTION. Terran Lane (University of New Mexico) presented a semi-supervised Bayesian approach to detecting intrusions on computers that subsumes traditional approaches of anomaly, which are unsupervised, and misuse detection, which are supervised.
11. ANOMALY DETECTION FOR VIDEO SURVEILLANCE. Robert Pless (Washington University in St. Louis) framed the problem of detecting anomalies in video surveillance data in terms of accurately identifying background regions that should be ignored. He used a probabilistic model to capture the local distribution of spatio-temporal image derivatives and showed that in many situations this information is sufficient to support anomaly detection.

The presentations covered an interesting variety of machine learning techniques that directly or indirectly build models of normal behavior and use this to detect abnormalities. Several common issues arose in the talks, including the need for developing explanations of the anomalies, dealing with domains that involve multiple time scales, and ensuring robustness to violations of modeling assumptions.

The participants also discussed strategies for improving the visibility of anomaly detection research within the greater communities of machine learning, data mining, and knowledge discovery. Proposed strategies included encouraging researchers to make extra effort to cite relevant technical work even if the domain of application is different from their own, targeting journals for special issues, and planning follow-up meetings on the topic. With respect to the last point, several attendees of the symposium are co-organizing a workshop on anomaly detection at the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in 2005. The proposed workshop has been accepted and its Web page can be found at <http://www.dmargineantu.net/AD-KDD05/>.

In summary, the Symposium on Machine Learning for Anomaly Detection encouraged the interchange of knowledge among researchers who have been addressing different applications that are nevertheless unified by their interest in learning models from data which can then be used to detect anomalies. Informal discussions during the meeting also suggested promising directions for future research in this increasingly important area of machine learning. Participants agreed that the symposium served important functions and expressed strong interest in continuing their interactions in additional meetings.