Data Mining Methods for Anomaly Detection KDD-2005 Workshop Report

Dragos Margineantu
The Boeing Company
Phantom Works
Math & Computing Tech.
P.O. Box 3707, M/S 7L-66
Seattle, WA 98124
dragos.d.margineantu@boeing.com

Stephen Bay
PricewaterhouseCoopers
Ten Almaden Boulevard
Suite 1600
San Jose, CA 95113
stephen.bay@us.pwc.com

Philip Chan
Florida Institute of Technology
Department of Computer Sciences
Melbourne, FL 32901
pkc@cs.fit.edu

Terran Lane
University of New Mexico
Department of Computer Science
Albuquerque, NM 87131
terran@cs.unm.edu

ABSTRACT

For many applications, data mining systems are required to detect anomalous (abnormal, unmodeled, or unexpected) observations. This has so far proven to be a difficult challenge because anomalies are usually considered to be "non-normal" observations, where "normality" is typically defined by very complex concepts. Because of these and other reasons, there are no standard and principled approaches for anomaly detection, yet, and the data mining processes that have led to successful solutions include most of the times ad-hoc (algorithmic, design, and implementation) decisions that incorporate prior or commonsense knowledge about the tasks that are addressed.

Consequently, we considered that it would be beneficial for both researchers and practitioners interested in anomaly detection and data mining, to organize workshop that would bring together people interested in this topic. We considered that the International Conference on Knowledge Discovery and Data Mining would be a good venue for such a workshop because of the diversity of interests, backgrounds, and problems that motivate people to attend the conference.

This paper describes the workshop on "Data Mining Methods for Anomaly Detection" – a one day event held in conjunction with KDD-2005 in Chicago, on August 21, 2005.

Keywords.

Anomalies, detection of anomalies in data, data mining, machine learning.

1. INTRODUCTION

The KDD-2005 workshop on "Data Mining Methods for Anomaly Detection" aimed to explore research efforts on data mining, machine learning, and related techniques that address the problem of detecting anomalies in data. The workshop was also an attempt to study the common tasks that need to be addressed in practical applications that require anomaly detection (AD) tools and algorithms such as data collection, sampling, and pre-processing.

Our main goals were to bring together researchers and practitioners interested in anomaly detection and to create a one-day forum for discussing recent advances in this area. We wanted to come to a better understanding of the practical challenges in developing and deploying data mining systems for anomaly detection and to inspire research on principled methods and techniques for detecting and predicting anomalies. Finally, we hoped to foster collaborations between AD researchers and those who had a need for AD – to bring together the data miners with the data stakeholders.

2. WORKSHOP OVERVIEW

"What is an Anomaly?" This was the first question the audience was asked during the opening remarks at the beginning of the day. The answers ranged from "three standard deviations away from the mean/normal" to "outliers" to "abnormalities that are different from outliers". The definitions proposed by the workshop organizers are:

- "*Anomalies* are occurrences of events that are unusual and that cannot be explained by our current knowledge of how the domain works. I see anomaly detection as simply trying to find these events. Anomalies are distinct from *outliers*, which I see as objects that on a measurement space, are far away from the other points." (Stephen Bay).
- "*Anomalies* are events that are not expected based on the *knowledge of previous events* that are considered normal. Since frequent events are usually considered normal, anomalies are usually rare. Based on the *normal* events, data mining methods can generate, in an automated manner, models (knowledge) of normalcy to identify *deviations*, which could be considered anomalies." (Philip Chan).
- "An *anomaly* is an event that *deviates* substantially from a known (explicit or implicit) *model* of some domain. *Anomalies* are events generated by a process that is *significantly different* than (explicitly or implicitly) *known/understood* processes. " (Terran Lane).
- "*Anomalies* are observations (or series of observations) with very *low likelihood of occurrence* with respect to (1) the *model(s)* that are believed (or are likely) to generate all observations and (2) the other observations that are available" (Dragos Margineantu).

These answers reflected the diverse interests of the people in the audience and the different motivations for their attending the workshop. We learned that most of the people in the audience were interested in this workshop because they were interested in addressing a particular application problem that involved

detecting anomalies, rather than in attempting to address a large or general class of different anomaly detection tasks. The submissions, the reviews, and our discussions prior to the workshop reflected that all of us - researchers and practitioners interested in data mining and learning methods for AD - typically have certain beliefs regarding the best approaches (algorithms, parameter settings, etc.) for a specific task. These beliefs are usually based on previous experience on other applications, data sets, and tasks. In practice, however, the practicalities of a real-world problem can dramatically change the approach we take.

The organizers' desire was to have a day of presentations and discussions on:

- Real problems encountered in anomaly detection;
- Practical solutions and their generalizability;
- The identification of small sub-tasks shared by different applications;
- Automating methodologies for anomaly detection;
- Techniques for assessing anomaly detection approaches;
- Identifying data sets that can be transformed into standardized testbeds for anomaly detection algorithms.

We hope that the presentations and discussions we had during this workshop will contribute to near-term steps in our community in addressing these issues.

3. WORKSHOP PRESENTATIONS

The workshop program included two invited talks and fourteen contributed presentations.

All materials related to the workshop are available from the workshop website: **http://www.dmargineantu.net/AD-KDD05**.

3.1 Invited Talks

3.1.1 Multi-Stage Classification

The invited speaker of the morning session of the workshop, was ***Ted Senator***, a program manager with the Information Processing Technology Office (IPTO) of the Defense Advanced Research Projects Agency of the United States (DARPA).

A common problem encountered in automated anomaly or rare-event detection is the large number of false alarms (or false positives) computed by the algorithms. Practice has shown that virtually all techniques that rely on a single stage approach (i.e., a single run of an algorithm) can reduce the number of false positives only at the (typically very high) expense of not detecting the anomalies. Based on this realization, our invited speaker presented his research on several approaches based on classification algorithms that reduce the number of false positive instances and in the meantime increase the true positive rate. The techniques presented in this talk were all based on two classification stages and were described by a unified architecture that allows different levels of complexity. The speaker presented experimental results on counter-terrorism data and on HIV-positive detection that show empirically that the proposed multi-stage classification methods exhibit improved accuracy especially in detecting extremely rare phenomena, and provide a reduction of false positives, over standard single stage approaches.

3.1.2 Outlier Detection in High-Dimensional Data – Using Exact Mapping to a Relative Distance Plane

The invited talk of the afternoon session was given by ***Ray Somorjai***, a scientist with the National Research Council of Canada (NRC), in Winnipeg. The focus in this presentation was the detection of outliers in highly-dimensional spaces, especially when the number of labeled observation is small. For the detection of the outliers, the speaker's approach is to employ different metrics for mapping the points into lower dimensional spaces and to employ a voting mechanism over the different mappings for the final decisions. The talk explored different mapping functions and analyzed their outlier detection capabilities on bio-medical data.

3.2 Contributed Presentations

3.2.1 Population-Wide Anomaly Detection

Weng-Keen Wong (the presenter) and his colleagues (Gregory Cooper, Denver Dash, John Levander, John Dowling, William Hogan, and Michael Wagner) have developed an algorithm – PANDA (Population-wide Anomaly Detection and Assessment) – designed to monitor health-care data and detect the inception of outbreaks caused by an outdoor, airborne release of inhalational anthrax. PANDA is basically a causal Bayesian network approach capable of incorporating multiple sources of data, domain knowledge, different types of evidence and, besides detecting the outbreaks, is capable of explaining the evidence that contributes most likely to its conclusions.

3.2.2 Strip Mining the Sky: The CTI-II Transit Telescope Survey

Until the beginning of last century, the term anomaly was used exclusively in describing abnormal movement of celestial objects. This talk (given by Peter Zimmer, and co-authored by John McGraw and their CTI-II computing collective colleagues from the University of New Mexico) described a large astronomical dataset that consists of images (approximately 200 gigapixels per night of operation) collected over a seven year period. Another database of parameters derived from the images, as well as extensive metadata, are stored. This presentation was a first step of a challenge coming from our colleagues working in astronomy. Having these databases available, will allow anomaly detection researchers to have a common ground for testing their methods and may help astronomers to discover interesting new space objects.

3.2.3 Learning to Live with False Alarms

Chris Drummond and Robert Holte addressed the problem of employing classification algorithms for highly imbalanced data (exhibiting class imbalances of 100:1, 1000:1 or higher). Two of the issues the practitioner has to address in these cases are that standard classification algorithms were not deigned to deal with severe data imbalance and that standard performance evaluation measures such as the misclassification rate are irrelevant. The authors further argue that the emphasis in dealing with high class imbalance should be on the costs/utilities of the decisions associated with the outputs of the classifier. If the end user is unhappy with the number of false alarms the only realistic answer

may be to demonstrate that cost calculations show that capturing a real event is worth any costs associated with the false positives.

3.2.4 Trajectory Boundary Modeling of Time Series for Anomaly Detection

Matthew Mahoney and Philip Chan addressed the problem of online detection of anomalous modes of mechanical failure, by employing only a small set of time series data from normal operation modes. The authors proposed two efficient techniques – path modeling and box modeling – that allow online scoring (during testing). These techniques were tested on two NASA shuttle valve datasets – TEK and VT1 (solenoid current and voltage measurements recorded on small valves that are used for actuating larger valves that control the fuel flow to space shuttle engines) against three other methods (Euclidean distance, dynamic time warping, and Gecko). The newly proposed methods outperformed the other methods on one of the datasets (TEK), and tied the performance of the other methods on the other domain (VT1). Path modeling has proven to be more accurate than box modeling, whereas the latter is significantly faster.

3.2.5 Provably Fast Algorithms for Anomaly Detection

In Don Hush's talk (joint work with Patrick Kelly, Clint Scovel, and Ingo Steinwart), he presented a summary of their ongoing work on the theoretical underpinnings of anomaly detection. In this talk, based on their recent JMLR paper on the subject, Hush formulated the anomaly detection problem in terms of minimizing an expected error criterion, *S(f)*, for a fixed AD classifier, *f*, defined as being the symmetric difference between the region classified as anomalous by *f* and the "true" anomalous region. The traditional difficulty with this formulation is that *S(f)* is impossible to measure directly from training data. In this work, Hush et al. demonstrate a risk function, *R(f)*, that upper bounds *S(f)* and that *can* be estimated from data. This surprising result reduces the anomaly detection problem to a traditional classification task and leads to the "DLD" (density level detection) support vector-based learning algorithm for AD. The authors gave empirical demonstrations that their algorithm yielded competitive or stronger performance on a cybersecurity anomaly detection domain, as compared to a number of previously published AD approaches.

3.2.6 Filtering Search Engine Spam based on an Anomaly Detection Approach

Kazumi Saito and Naonori Ueda address the problem of detecting search engine spam, which are websites that contain no relevant information but score highly on a search engine's ranking solely because of a densely connected link structure. They attempt to detect spam websites by finding network cores which is a set of sites whose connections are anomalously dense. They take an iterative approach by computing the eigenvectors of the network adjacency matrix, identifying an anomalously dense subset, removing it, and repeating the process.

3.2.7 Discovering Hidden Association Rules

The joint work of Marco-Antonio Balderas, Fernando Berzal, Juan-Carlos Cubero, Eduardo Eisman, and Nicolás Marín on "Discovering Hidden Association Rules", addressed anomaly detection in the context of the traditional KDD question of association rule mining.

As has been widely noted in the data mining literature, association rule mining algorithms tend to produce a large number of often redundant rules. In the AD context, however, we are interested not in rules that describe a large fraction of the data, but in rules that describe interesting out-of-the-norm cases. Further, such cases are often *infrequent*.

This group proposed the definition that an anomalous association rule is a rule that is confident only in the *absence* of a rule that is both confident and frequent. For example, if a confident and frequent rule is "if symptom-X then disease-Y", then an anomalous case might appear when the consequent doesn't hold: "if symptom-X and _not_ disease-Y then disease-A". Such rules are constrained to be confident, but need not have strong support, as we are interested in possibly rare anomalies. The authors demonstrated the ATBAR (Anomaly TBAR) algorithm: a rule-mining process that can extract such relations. They showed the performance of ATBAR on a variety of medical classification problems drawn from the UCI database, demonstrating that it was able to extract compact anomaly rule sets from moderately large data sets.

3.2.8 Multivariate Dependence among Extremes, Abrupt Change and Anomalies in Space and Time for Climate Applications

Auroop Ganguly's talk (joint work with Tailen Hsing, Rick Katz, David Erickson, George Ostrouchov Thomas Wilbanks, and Noel Cressie) discussed multivariate spatio-temporal dependencies between extreme and unusual values, as well as sudden changes in climate data. The authors have provided an overview of current approaches and of technical challenges for the data mining community. They also presented their work on visualization and quantification of multivariate dependences between anomalous values.

3.2.9 Anomalous Spatial Cluster Detection

Daniel Neill (the presenter) and Andrew Moore have given the audience a detailed overview of their *generalized spatial scan* framework for the task of detecting spatial clusters. This task involves identifying locations, shapes, and sizes of potentially anomalous regions of points in space, and discriminating between real clusters and chance occurrences of points in a region.

The steps of the generalized spatial scan are described below and include

- Gathering data for a set of spatial locations. The goal is to find the regions for which the counts are higher than expected (given a the baselines). Population-based and expectation-based methods are the two typical classes of approaches for this task.
- Choosing the set of spatial regions to search over. The crux is to select regions that are partially overlapping, cover the entire space, and the number of selected regions should not be too large (because of computational infeasibility) nor too little (which would result in reduced power).
- Choosing models of the data under the null hypothesis of no clusters – H_0, and the alternative hypothesis assuming a cluster in region S (from the selected regions) – $H_1(S)$.

- Deriving a score function F(S) based on H_0 and $H_1(S)$.
- Finding the regions with the highest score values.
- Filtering the high-scoring regions to select the most "interesting" ones. A frequentist approach (using randomization testing for calculating the statistical significance) and a Bayesian approach (relying on the posteriors of each potentially interesting cluster) have been presented.

3.2.10 Current and Potential Statistical Methods for Anomaly Detection in Modern Time Series Data: The case of BioSurvellance

Galit Shmueli takes a retrospective look at methods for biosurveillance which is a field concerned with the early detection of disease outbreaks, both natural and as a result of a terrorist attacks, by monitoring syndromic data. She discusses the current practice which is based on traditional statistical methods such as statistical process control and autoregressive time series models but notes that these methods depend on strong assumptions which are almost always false. Consequently, she examines similar problems in related fields, surveys their methods, and discusses whether the field of biosurveillence can borrow techniques from their work.

3.2.11 An Empirical Comparison of Outlier Detection Algorithms

Matthew Otey, Srinivasan Parthasarathy, and Amol Ghoting presented an empirical comparison of three outlier detection methods on an intrusion detection data set. The three methods they compared were Orca, a distance based outlier detection method, LOADED, a method based on computing an itemset lattice, and RELOADED, a method based on using classifiers to model dependencies between features. They examine the quality of the results, execution time, and memory requirements.

3.2.12 Detecting Anomalous Patterns in Pharmacy Retail Data

Maheshkumar Sabhnani has presented his joint work with Daniel Neill and Andrew Moore on a bio-surveillance system they developed for daily monitoring of over-the-counter pharmacy sales and detecting anomalous patterns that may be indicators of disease outbreaks. Complex trends (seasonal, day-of-week, etc.) in the pharmacy sales data, missing data, partial lack of labels, and the high costs of false positives are only some of the challenges in developing an automated data mining approach for this task. The presented approach relies on fast space-time scan statistics which is capable of incorporating expert domain knowledge. The authors have described how the system has evolved over time, how different technologies have been incrementally added, and how these techniques influenced the accuracy of detecting online the real anomalies in the data collected by a monitoring system (The National Retail Data Monitor) of the pharmacy data, operated at the University of Pittsburgh's Real-time Outbreak and Disease Surveillance Laboratory.

3.2.13 A Comparison of Generalizability for Anomaly Detection

Gilbert Peterson, Robert Mills, Brent McBride, and Wesley Allred investigate the issue of how tightly an anomaly detection system should model behavior of normal examples. Specifically, in an intrusion detection system one could argue that the model of normal class behavior should fit the data sample as tightly as possible and only cover behaviors that were actually observed since attacks might try to mimic the normal class. They investigate this hypothesis by comparing three different approaches for modeling normal data: k-means with spheres, k-means with ellipsoids, and convex polytopes. They discover that although the convex polytope method develops models that are most specific to the normal class, it can often perform worse than the k-means approach because of generalization issues.

3.2.14 An Empirical Bayes Approach to Detect Anomalies in Dynamic Multidimensional Arrays

Deepak Agarwal addressed the problem of detecting anomalies in cross-classified data streams in which each dimension corresponds to different levels of a categorical variable. The actual task motivating the approach is a speech mining task: to automatically extract important service and business intelligence information from records of dialogs of customers calling an automated help desk. The problem addressed in this talk was the detection of changes in the features selected from dialogues. The proposed hierarchical Bayesian approach – *hbmix*, a method that works by adjusting for marginal effects – was compared against a naïve multiple comparisons, per-comparison error rate (PCER) technique.

4. RESOURCES AND CONTACTS

One goal of our workshop was to put "data stakeholders" in touch with anomaly detection researchers. We had interest from a number of such groups who made long or short presentations to the workshop:

- Peter Zimmer's long presentation (Section 3.2.2) closed with a plea to the audience: The astronomers are overwhelmed by their data and critically need assistance in analyzing it and locating interesting anomalies from it. They are very interested in building collaborations with anomaly detection and KDD researchers. Dr. Zimmer can be reached at zimm@as.unm.edu.
- Drs. Dennis Glanzman (National Institute of Mental Health) and Yuan Liu (National Institute of Neurological Disorders and Stroke) from the National Institutes of Health (NIH) addressed the audience briefly. There are a myriad of problems in the analysis of neural and neuroimaging data that could greatly benefit from anomaly detection and KDD techniques. Dr. Glanzman can be contacted at glanzman@helix.nih.gov, while Dr. Liu can be reached at yl5o@nih.gov.
- Kendra Moore at DARPA's IXO office is initiating a program on anomaly detection for analysis of naval deployment activities (the PANDA program). For more information on this program, please refer to the program solicitation at http://dtsn.darpa.mil/ixo/solicitations/panda/index.htm.

5. DISCUSSION

At the end of the day, the organizers have wrapped up the one-day workshop by identifying some directions of future research. These include:

- Capture and representation of rich knowledge, and the incorporation of rich domain knowledge in automated data mining processes for anomaly detection;
- Reliable anomaly detection components for large scale automated decision systems;
- User-customized anomaly detection systems (i.e., systems that can be customized and used by people who are not data mining or machine learning experts);
- What is the equivalent of overfitting for anomaly detection tasks, and what are principled approaches to address it?
- Principled validation and testing techniques for anomaly detection algorithms and tools: (1) statistical tests, (2) the need for testbeds for anomaly detection systems;
- Principled feature construction methods for AD?
- Anomaly detection techniques for adversarial tasks.

6. WORKSHOP PROGRAM COMMITTEE

The program committee members have reviewed papers and have helped the organizers in putting together the workshop program. Each submission to the workshop has been reviewed by at least two program committee members. Our workshop program committee was composed of:

Naoki Abe - IBM TJ Watson
Carla Brodley - Tufts University
Vincent Clark - University of New Mexico
Diane Cook - University of Texas, Arlington
Chris Drummond - The National Research Council of Canada
Wei Fan - IBM TJ Watson
Roman Fresnedo - The Boeing Company
Eamonn Keogh - University of California, Riverside
Adam Kowalczyk - National ICT Australia
Aleksandar Lazarević - University of Minnesota
Wenke Lee - Gerogia Institute of Technology
John McGraw - University of New Mexico
Ion Muslea - Language Weaver, Inc.
Raymond Ng - University of British Columbia
Galit Schmueli - University of Maryland, College Park
Mark Schwabacher - NASA, Ames Research Center
Salvatore Stolfo - Columbia University
Weng-Keen Wong - Oregon State University
Bianca Zadrozny - IBM TJ Watson

7. ACKNOWLEDGMENTS

We would like to thank to all presenters, attendees, and program committee members – they are the main contributors to the success of the workshop.

We would also like to extend our thanks to the KDD-2005 Workshop Chair - Mohammed Zaki, to the local chairs – Shirley Connelly, Bamshad Mobasher, and Peter Caron, to the proceedings chair – Jaideep Vaidya, to Bing Liu, and to all KDD-2005 conference organizers.

The workshop has been sponsored by The Boeing Company and by PricewaterhouseCoopers, Inc. Their financial support has enabled the participation of some of the presenters.

About the authors:

Dragos Margineantu is a Computer Scientist with the Adaptive Systems group of Boeing's Mathematics and Computing Technology organization. His research interests include learning and decision systems for detecting anomalies, integration of domain knowledge into learning and decision processes, learning and data mining in adversarial environments, software engineering for learning and decision systems, cost-sensitive and active learning. At Boeing, Dragos Margineantu has developed the learning components of software tools for maintenance operations, manufacturing process optimization, and security applications. He currently manages research projects on Validation and Testing of Decision Systems and on Learning Systems. Dragos Margineantu earned his Ph.D. in Computer Science from Oregon State University. (http://www.dmargineantu.net)

Stephen Bay is a Scientist with the Center for Advanced Research at PricewaterhouseCoopers. His current work focuses on detecting fraud in financial data using advanced data mining and statistical methods. He earned his Ph.D. in Information and Computer Science from the University of California, Irvine. (http://www.isle.org/~sbay)

Philip Chan is an Associate Professor of Computer Sciences at the Florida Institute of Technology. His research interests include intrusion detection, scalable data mining and machine learning methods, and web personalization. He has developed anomaly detection techniques for computer security and device monitoring tasks. Professor Chan earned his Ph.D. in the Computer Science at Columbia University. (http://www.cs.fit.edu/~pkc)

Terran Lane is an Assistant Professor of Computer Science at the University of New Mexico. His interests vary across the field of machine learning, but include topics such as anomaly detection, reinforcement learning, decision making, Bayesian networks and Bayesian data modeling, unsupervised learning, and relational and graph learning. He is particularly interested in application-driven learning problems, and his work has centered around practical applications such as computer security, robotics, user modeling, bioinformatics, and neuroinformatics. Professor Lane earned his Ph.D. in the Department of Electrical and Computer Engineering at Purdue University and spent two years as a Postdoctoral researcher at MIT before joining the faculty at UNM in 2002. (http://www.cs.unm.edu/~terran)