An ODE (Outlier Detection and Explanation) to Scalable Anomaly Detection: Algorithms and Applications

Srinivasan Parthasarathy
Ohio State University
srini@cse.ohio-state.edu

Funding Acknowledgements: NSF, DOD, NIH, Intel, Pfizer.
The Data Deluge

“Every 2 days we create as much information as we did up to 2003”

- Eric Schmidt, Google ex-CEO
Data Storage Costs are Low

600$ to buy a disk drive that can store all of the world’s music

[McKinsey Global Institute Special Report, June ’11]
All this data is only useful if we can *scalably extract* actionable knowledge
Challenges

1. Scalability

*Dictated by real time constraints and timeliness*
Challenges

2. Noise and Uncertainty

*Inherent to many real-world problems*
Challenges

3. Domain Specific Needs and Guidance

Need to bridge impedance mismatch
Challenges

4. Dynamics and Novel Structure

Need for models that can accommodate
5. Cognitive Overload

Need to support guided interaction for human in the loop
Outlier Definition [Hawkins’80]

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”
Contemporary Methods

- Geometric
- Statistical
- Density-based
- Distance-based
- Sub-space methods
- Localized methods

Different types of data and contexts in which they arise
- Images
- Text
- Numeric/quantitative
- Categorical
- Streaming

See number of recent surveys and tutorials on these topics
Scalability

Focus on Distance-based Outlier Detection

Joint work with Y. Wang, W. Meira, C. Teixeira
Distance Based Outlier Detection

1. Outliers are objects with fewer than \( k \) neighbors in the database, where a neighbor is an object within distance \[Kn'99\].

2. Outliers are the \( n \) objects presenting the highest distance values to their respective \( k \)th nearest neighbor (kNN) \[RRS'00\].

3. Outliers are the \( n \) objects presenting the highest average distance to their respective \( k \) nearest neighbors \[AP'02\].
Standard NL Algorithm

- Naïve: For each point x in dataset
  - y = Find kth nearest neighbor to point x
  - Outlier score OS(x) = dist (x,y)
  - Pick top L points with largest outlier score

- Basic Optimizations
  - Always maintain top L outliers and the outlier score of worst outlier z (min(L-score) = OS(z))
  - If for point x after processing all neighbors OS(x) > OS(z)
    - Add x to L; remove z; update L-score;
  - Early pruning (ANNS) – if point x has seen more than k neighbors and its current outlier score OS(x) < OS(z); x can be pruned.
Advanced Optimizations (used by many)

A. **ROCO** – Ranking Object Candidates for Outliers
   - Can enable downstream pruning of points through ANNS and often leverages pre-partitioning or randomization [ORCA, RBRP]

B. **ROCN** – Ranking Object Candidates for Neighbors
   - Can potentially process point X faster and typically leverages pre-partitioning and prior distribution [WJ’07]

C. **PPSO** – Pruning Partitions during Search for Outliers
   - Can eliminate entire partitions based on summary statistics/pre-partitioning [RRS’00]

D. **PPSN** – Pruning Partitions during Search for Neighbors
   - Can eliminate entire partitions during neighborhood search leveraging pre-partitioning(index trees, VA-files) [RRS’00]
   - Used in strawman by RRS – largely ignored by follow up work
Factorial Design Experiment [VLDB’10]

- To understand interactions among optimizations:
  - Baseline – ANNS; pre-partitioning
  - Factors: A ROCO; B ROCN; C PPSO; D PPSN

- Insights
  - Using all (ABCD) or all except C (ABD) most effective
  - A & D are dominant factors
  - No single factor dominates
  - Factor interactions often data dependent
Lessons Learned and Followup

- Important to understand interactions among various optimizations
- Also important to understand interactions that affect quality
- E.g. Interactions between pre-processing and predictive analytics
- Interesting that a strawman factor (D. PPSN) largely ignored by community turned out to be highly relevant

- Investigated more refined ROCO strategy
  - Driven by Locality Sensitive Hashing [IM’99]
  - Order of magnitude improvement on IDS, and Climate datasets [ICDE’11; ICDE’13]
Leveraging Domain Guidance

Understanding Defect/Crack Propagation in Materials

Joint work with J. Wilkins, R. Machiraju, and a cast of many
Tracking Defects in Materials

- **What?**
  - Understanding formation and propagation of defect structures

- **Why?**
  - Fundamental problem in semiconductor industry (Bo-Si) etc.
  - Impacts yield rates

- **How?**
  - By analyzing data from large physics-driven simulations (typically MD/QM).
Context: Role of Outlier Detection

- Defects are defined as anomalous structures in regular lattices

- Challenges: In-vivo; data-scale; velocity, uncertainty (jitter)

- Key Insight: Physical laws can guide feature selection and search space → need to incorporate
  - E.g. Bond angles; Bond lengths; Neighborhood features

- Spatio-Temporal Approach
  - Detect and group anomalies → defect characterization → track over time and space → understanding crack formation
Results

- Efficiency
  - Able to analyze at better than frame rate for previously seen structures
  - Created catalogue/index of defects and their evolution
  - Facilitated in-vivo computational steering around novel defects

- Accuracy > 99%
  - Ability to withstand thermal noise + Rotational and translational invariance
  - Incorporating guidance from physics-based models was crucial

![Bar chart showing time in hours for different datasets]
Human Factors and Providing Visual Explanations

Disease Diagnosis: Keratoconus & Glaucoma

Joint work with M. Twa, K. Marsolo & M. Bullimore
Case Study 2: Keratoconus Diagnosis

- Progressive, degenerative, non-inflammatory disease.
- Blindness and Corneal Transplant

Diagnosis Procedure
- Video-keratography & Manual analysis

Challenges to early detection
- Voluminous
- High dimensional data
- Features of interest small in scale to mean shape – localized anomalies

Late stage Keratoconus  Normal (clinically ideal)
Modeling with Zernike Polynomials

- Hyper-geometric radial basis functions
  - Orthogonal building blocks
  - Lower order → basic shape
  - Higher order → harmonics
- Compact representation
- Anatomic correspondence to clinical concepts
Contrasting Anomalies with Decision Surfaces

- **Decision Support Objectives**
  - **Guide** clinician on why patients are classified the way they are.
  - **Contrast** an individual patient with others in the same group

- **How?**
  - Zernike Fit of Corneal Shape
  - Coefficients/Modes – specific features of the cornea.
  - Can use as building blocks to construct a surface and contrast
Visual Explanations of Outliers

Rule 1 - Keratoconus

Rule 4 - LASIK

Rule 8 - Normal
Dynamics and Noisy Data

Pathodological Dynamics of Pharamaceutical Trials: Hepatotoxicity

Joint work with M. Otey, C. Trost, P. March & A. Friedman and a cast of many
Analyzing Clinical Trials Data

Why?
- Study Efficacy
- Examine Hepatotoxicity (e.g. liver panel)
  - e.g. Vioxx/NSAID

How?
- Qualitative, involving clinical judgement
- Ad-hoc metrics – e.g. Hy’s rule – univariate approach

Challenges
- Irregularly spaced time series; individual variability; uncertainty; missing data; stochastic covariate structure
Approach

- Data preprocessing (critical)
  - Logarithmic transformation + centralization

- Stochastic analysis of covariance of liver analyte panel
  - Patients on placebo vs Patients on drug
  - Q-Q plots of both groups indistinguishable

- Extract distance based anomalies
  - Compare against expectation – hypothesis testing
**Exemplar results**

- Top right – drug with no observable toxic effects.
  - In market

- Bottom right – a greater than expected number of anomalies are triggered from those on the drug – possible hepatotoxicity
  - Early detection – 2 years before drug testing was discontinued
Novel Structure

Medical and Emergency Response

Joint work with A. Sheth, Y. Ruan, D. Fuhr, H. Purohit and a cast of many
Complex Data

- Physical Sensing
- Simulated Storm Surge Modeling
- Citizen Sensing (Tweets)
- Underpinning network structure
  - Cellphone
  - Social

Goal: Extract actionable summaries that can help prioritize and coordinate relief activities for emergency response personnel
  - Outliers are often key actionable elements
Spatial Filtering by Storm Surge Models
Thematic Filtering

- Spatial themes
- Topical themes
  - “Help”, “Danger”, “Donate”
- Temporal themes
  - Before/During/After Disaster
Key Findings

- Spatial Thematic Filtering
  - Prioritization of Critical Infrastructure and People Needs to extract Contrast Patterns
  - “Frankenstorm” vs. legitimate calls for help

- Topic-Temporal Thematic Filtering
  - Highlights Variance in Influence, Passivity, and Trustworthiness
  - Finds Contextual Outliers (e.g. MOVEON.org)
Outlier Detection is a crucial element of the data science toolbox. Understanding its interactions with other elements is crucial.

Methodological challenges:
- Scalability, Complex data, Understanding interactions, Noise, Dynamics
- Design of Experiments (factorial experiments) can shed important insights

Human Factors (guidance and offering explanations) is key.

Lots of cool applications:
- Analyzing Games for Therapy (coming soon)
Thanks and the promised ODE!

Upon us is the complex data deluge,
in scalable analytics we must take refuge.
Guidance and explanations from the domain matter,
to put actionable outliers and anomalies on a platter.

- Organizers
- Collaborators
- Audience
- Funding Agencies
References

Edwin M. Knorr, Raymond T. Ng: Algorithms for Mining Distance-Based Outliers in Large Datasets. VLDB 1998: 392

Sridhar Ramaswamy, Rajeev Rastogi, Kyuseok Shim: Efficient Algorithms for Mining Outliers from Large Data Sets. SIGMOD Conference 2000: 427-43

Fabrizio Angiulli, Clara Pizzuti: Fast Outlier Detection in High Dimensional Spaces. PKDD 2002: 15

Stephen Bay, Mark Schwabacher: Mining distance-based outliers in near linear time with randomization and a simple pruning rule. KDD 2003: 29


Amol Ghoting, Srinivasan Parthasarathy, Matthew Eric Otey: Fast Mining of Distance-Based Outliers in High Dimensional Datasets. SDM 2006: 609-613


Matthew Eric Otey, Srinivasan Parthasarathy, Donald C. Trost: Dissimilarity Measures for Detecting Hepatotoxicity in Clinical Trial Data. SDM 2006: 509-516

Mahashweta Das, Srinivasan Parthasarathy: Anomaly detection and spatio-temporal analysis of global climate system. KDD Workshop on Knowledge Discovery from Sensor Data 2009: 142-150


Heman Purohit, Yive Ruan, David Fuhr, Srinivasan Parthasarathy, Amit P. Sheth: On Understanding the Divergence of Online Social Group Discussion. ICWSM 2014


Michael D. Twa, Srinivasan Parthasarathy, Thomas W. Raasch, Mark Bullimore: Decision Tree Classification of Spatial Data Patterns from Videokeratography using Zernicke Polynomials. SDM 2003: 3-12