Graph-Based Anomaly Detection with Soft Harmonic Functions

Michal Valko

Advisor: Milos Hauskrecht
Anomaly (Outlier) Detection

- **Goal:** identify unusual patterns in data

- **Focus:** conditional anomalies

- **Contribution:** graph-based method for conditional anomaly detection

- **Application:** medical error detection
Patient electronic records have: demographics, conditions, labs, medications administered, procedures performed,...
**Conditional Anomaly**

Assumption: *Conditional* anomalies correspond to medical errors.

“Medical overspending for 200,000 preventable deaths”

(HealthGrades study, Wall Street Journal, July 27th 2004)

“Controlling overspending becoming necessary”

Dr. Reinhardt

(Health Care Costs, New York Times, December 19th 2010)
Traditional Anomaly Detection

- Nearest Neighbor
  - Distance – anomalies are distant (NN)
  - Density – anomalies in low density regions (LOF, COF, LOCI)

- Classification
  - Model based (separate models for (ab)normal distributions)
  - 1-class (1-class SVM)
  - Classify normal vs. abnormal (when labels available)

- Statistical
  - > 3std
Challenges for CAD

Task: detect anomalies in labels

Dataset adopted from [Papadimitriou and Faloutsos, 2003]
Related Work (CAD)

- Cross Outlier Detection (Papadimitriou, 2003)
CAD approaches

Class Outlier Approach

- OneClass SVM, LOF, ...

Discriminative Approach

- SVM-CAD

Regularized Discriminative Approach

- Connectivity AD, Soft Harmonic AD

regularizing unconditional outliers
Conditional Anomaly Detection Goal

Problem statement (★): For a dataset \((x_i, y_i)_{i=1}^{n}\) find pairs of \((x_i, y_i)\) such that \(P(y \neq y_i | x_i)\) is high.

... and at the same time avoid unwanted anomalies

both train and test data are labeled
Class Outlier Approach

- Take a test case \((x, y)\)
- Take any unconditional anomaly method
- Find out if \(x\) is anomalous with respect to \(\{x | x \text{ has class } y\}\)

**Problems:**
- Fringe points
- Unconditional outliers
- Anomaly (alert) scores for class 1 and class 2 may not be comparable
Discriminative Approach

- \( P(y'|x) \) is high → conditional anomaly
- Learn Model/Build Projections
- Bayes Network

\[ d(y|x) = P(y'|x) \quad y' \neq y \]

- bigger the alert score → more anomalous

- Support Vector Machines projections

\[ d(y|x) = -y(w^T x + w_0) \]
Support Vector Machines projections

More Anomalous                                   Less Anomalous

5    ………..    1         0
-1   ………
-5

\[ d(y|x) = -y(w^T x + w_0) \]

[Valko et al., 2008]
[Valko and Hauskrecht, 2008]
[Hauskrecht et al., 2010]
A new approach

- **Disadvantages of the SVM-CAD**
  - only linear decision boundary
  - can become overly confident in the areas with little data
    - Isolated points (unconditional outliers)

- **Soft Harmonic Anomaly Detection**
  - Non-parametric
  - Graph-based
  - Regularization
    - Control the influence of unconditional outliers
  - Can incorporate **unlabeled** examples
    - Missing medical records
    - Tests not done frequently because of the budget constraints
Dealing with Outliers

- apple
  - -0.9
  - -1
  - -0.9
  - -0.8
- kiwi – outlier
  - -0.7
  - -0.4
  - -0.2
- orange
  - 0.6
  - 0.6
  - 0.8
  - 1

SINK
0
Dealing with Outliers

apple

orange

kiwi – outlier
Regularization

Low confidence labeled data

\( \gamma_\text{g} = 100.000 \)

\( \gamma_\text{g} = 1.000 \)

\( \gamma_\text{g} = 0.010 \)

ONLY LABELED

ALL DATA
Soft Harmonic Solution

- Unconstrained Regularization

\[
\min_{\ell \in \mathbb{R}^n} (\ell - y)^T C (\ell - y) + \ell^T K \ell
\]

- Close form solution

\[
\ell = (C^{-1} K + I)^{-1} y
\]

- when \( \ell_i \) is rewritten as \( |\ell_i| \ \text{sgn}(\ell_i) \)

- \( |\ell_i| \) can be interpreted as a confidence

- \( |\ell_i| >> 0.5 \) and \( \text{sgn}(\ell_i) \neq y_i \) Conditional Anomaly!
Synthetic Data

- evaluation of conditional anomaly methods is challenging
- synthetic data with known distribution
- flip 3% of the labels
- compare how the anomaly score agrees with true score
**Synthetic Data: Results**

- **Evaluation metric:**
  - How the anomaly score agrees with the true score

<table>
<thead>
<tr>
<th></th>
<th>Dataset D1</th>
<th>Dataset D2</th>
<th>Dataset D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>QDA</td>
<td>73.8% (2.1)</td>
<td>29.4% (5.2)</td>
<td>61.0% (1.2)</td>
</tr>
<tr>
<td>SVM</td>
<td>58.8% (7.0)</td>
<td>49.8% (1.7)</td>
<td>46.1% (3.1)</td>
</tr>
<tr>
<td>1-class SVM</td>
<td>51.3% (0.9)</td>
<td>47.7% (0.6)</td>
<td>64.7% (0.7)</td>
</tr>
<tr>
<td>wk–NN</td>
<td>74.2% (1.9)</td>
<td>56.5% (1.7)</td>
<td>61.4% (2.1)</td>
</tr>
</tbody>
</table>

**Better**
Top 5 best scoring anomalies for different methods on the synthetic dataset D3

- True Model
- Quadratic Discriminant Analysis
- SVM with RBF kernel
- 1-class SVM with RBF kernel
- Soft Harmonic Anomaly Detection
- weighted k-NN

Our method: TRUE

True

Method: OUR METHOD
Medical Data

- 4486 patients from UPMC
- Cardiac surgery (2002-2007)
- 45767 patient-day events/states
- 9K attributes
- 222 states evaluated by 15 experts

- nearest neighbor graph
- Metric: How much the score agrees with the experts.

1. Laboratory tests (LABs)
2. Medications (MEDs)
3. Visit features/demographics
4. Procedures
5. Heart support devices
PCP data set: Segmentation
PCP Dataset: PLT Lab feature

- Last value: A
- Last value difference = B-A
- Last percentage change = (B-A)/B
- Last slope = (B-A) / (tB-tA)
- Nadir = D

Nadir difference = A-D
Nadir percentage difference = (A-D)/D
Baseline = F
Drop from baseline = F-A
Medical Data Results

• Outperforming SVM method over the range of settings of regularization parameters
Medical Data Results

- Outperforming standard weighted nearest neighbors on the same graph

![Graph showing AUC comparison between Soft Harmonic CAD and CAD with weighted k-NN](image-url)
Conclusion & Future Work

- A non-parametric graph-based approach
  - Successfully detect conditional anomalies

- Future work
  - Online Soft Harmonic Anomaly Detection
  - Parallelization of harmonic solution

Adapt to changes in medical practice

Thanks to:
Branislav Kveton, Greg Cooper, Tomas Singliar, Shyam Visweswaram, Iyad Batal, Amy Seybert, Hamed Valizadegan, Saeed Amizadeh, Quang Nguyen, Dave Krebs