ASK-the-Expert: Active learning based knowledge discovery using the expert

Kamalika Das
Data Science Group
NASA Ames Research Center

ML Workshop, August 2017
Roadmap

• Problem description
• State-of-the-art
• Proposed framework
• Tool description
• Algorithms
• Performance analysis
• Summary
Problem

- Identify safety events in flight operational data
- Unsupervised anomaly detection
- SME review of anomalies
Unsupervised anomaly detection

- Lack of definition of ‘safety’ incident
- One-class SVM based anomaly detection

State of the art

Data Collection
- ATRCC
- TRACON
- FAA Facilities

Data Processing
- ARTCC
- TRACON

Data Merge

Existing System

Operationally Significant Events
- Labels
- Anomalies
- Nominals
- MKAD: Unsupervised Anomaly Detection

Feature Selection And Normalization

Calculate Flight Separation and Turn-to-final features
Proposed approach

Active learning with rationales framework
Active learning framework

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$f^*$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$f^*$</td>
</tr>
<tr>
<td>$f_3$</td>
<td>$f^*$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$f_n$</td>
<td>$f^*$</td>
</tr>
</tbody>
</table>

**Statistical anomalies**

**Flight**: $x^*$  
**Label**: $y^*$  
**Rationale**: OS  
Loss of separation

**Labeled pool**

**Unlabeled pool**

**Active Learner**

Model: 2-class multiple kernel SVM
Active learning strategy: Most likely positive
Automated feature construction: Multiple kernel learning + decision tree construction

Sample to label
ASK-the-Expert tool: architecture

Diagram showing the architecture of the ASK-the-Expert tool.

- **Active Learner**
- **Database**
- **Coordinator**
- **Annotator**
- **Investigator**

The diagram illustrates the flow of data and instances:
- Data packets feed into the Coordinator.
- The Coordinator sends instances to label to the Investigator.
- The Investigator labels instances and sends them back to the Coordinator.
- The Coordinator then sends labeled instances to the Annotator.
- The Annotator sends a ranked list of anomalies back to the Coordinator.

Key processes:
- Unsupervised anomaly detection
- Instances to label
- Ranked list of anomalies
Annotator component
Coordinator component

Ingest annotations from SME
- Use text mining based mapping from annotations to features
  - Latent dirichlet allocation (LDA)
  - Neural networks

Learn weights of most important features
- Simple MKL

Automated feature construction
- Discretization of time series (SAX)
- Decision tree induction

Classifier learning
- 2-class multiple kernel Support Vector Machine
Multiple kernel support vector machine

- Multiple kernel 2 class SVM: classifying between operationally significant (OS) and uninteresting (NOS) flights

<table>
<thead>
<tr>
<th>Flight time series</th>
<th>Feature set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>f₁ f₂ f₂ ... ... fₙ</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td></td>
</tr>
</tbody>
</table>

Đecision function:

\[ f(x) = \sum_{i} \alpha_i K(x_i, x) + b \]

2-class SVM objective:

\[ \min_{\alpha} D(\alpha) = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \Phi(x_i) \cdot \Phi(x_j) - \sum_{i} y_i \alpha_i \quad \text{s.t.} \left\{ \begin{array}{l} \sum_{i} \alpha_i = 0 \\ 0 \leq y_i \alpha_i \leq C \end{array} \right. \]
Rationale feature construction

- How to set weights: $\eta_1, \eta_2, \ldots, \eta_n$
  \[
  K_\eta = \sum_{m=1}^{P} \eta_m k_m(x_i^m, x_j^m) \quad s.t. \eta_m \geq 0 \& \sum \eta_m = 1
  \]

- Simple MKL algorithm
  - Modified objective function
  - Alternates between optimizing classifier margin and weights of kernels
Rationale feature construction

• Decision tree induction

```
SpeedOfIntercept <= -1.3962
  gini = 0.0624
  samples = 31
  value = [30, 1]
```

```
Latitude_mean <= 0.7325
  gini = 0.5
  samples = 2
  value = [1, 1]
```

```
MaximumOvershoot <= 3.7731
  gini = 0.0624
  samples = 31
  value = [30, 1]
```

```
True
```

```
Latitude_mean <= 0.7325
  gini = 0.5
  samples = 2
  value = [1, 1]
```

```
False
```

```
  gini = 0.0
  samples = 29
  value = [29, 0]
```

```
  gini = 0.0
  samples = 1
  value = [0, 1]
```

```
True
```

```
  gini = 0.0
  samples = 30
  value = [30, 0]
```

```
False
```

```
  gini = 0.0
  samples = 1
  value = [0, 1]
```
Data

ORIGINAL FEATURES
- Latitude
- Longitude
- Altitude
- Ground speed
- Horizontal separation
- Vertical separation
- Aircraft size
- Turn-to-final (TTF) parameters:
  - Maximum overshoot
  - Speed at TTF
  - Distance at TTF
  - Angle at TTF
  - Altitude difference at TTF
- Nearest neighboring (NN) flight info:
  - NN flight on same runway
  - NN flight on parallel runway
  - NN flight part of the same flow
“Loss of separation”
- Horizontal separation < 3 miles AND Vertical separation < 1000 ft AND nearest neighboring flight is not on parallel runways and not part of the same flow

“Large overshoot”
- Maximum overshoot is greater than a threshold based on values of flights with positive labels

“Unusual flight path”
- Overall deviation from expected (average) trajectory of all landing flights on that runway
Experimental setup

• Data set: 30 NM airspace around Denver International Airport for Aug 2014
  – Training set: ~2400 flights
  – Statistical anomalies: 153
  – OS flights: 24

• 2 fold cross validation with 10 random bootstraps for each fold
Performance analysis

• Metrics: precision@5 and precision@10
• Most-likely positive strategy

\[ x^* = \arg \max_{x \in \mathcal{U}} P_\theta(\hat{y}^+ | x) \]

Learning curves for different active learning strategies
Performance analysis

Learning curves for most likely positive strategy with and without rationales

75% savings in labeling effort
Performance analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Target \textit{precision@5}</th>
<th>Target \textit{precision@10}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>RND</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>MKAD-Sampling</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>MLP</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>MLP_w/Rationales</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Comparison of number of labeled flights required by various strategies to achieve a target performance measure. ‘n/a’ represents that the target performance cannot be achieved by a method even with 45 labeled flights.
Performance benefits

• Generalization
  – Two different test data sets: July 2014 and July 2015
  – Average improvement in precision@5: ~30%
  – Average improvement in precision @10: ~65%

• Review time
  – Up to 75% reduction in review time for same target performance
Summary

• Goal: to reduce SME review time of statistical anomalies identified using unsupervised anomaly detection

• Use active learning with rationales to learn 2-class classifier to distinguish between operationally significant and uninteresting anomalies

• Classifier generalizes to other data sets from the same domain

• Up to 75% reduction in SME review time
Acknowledgement

• This work is supported by Center Innovation Fund (CIF) 2017 award

• Team:
  – Nikunj Oza, NASA Ames Research Center
  – Bryan Matthews, SGT Inc.
  – Illya Avrekh, SGT Inc.
  – Manali Sharma, PhD Student, Illinois Institute of Technology
  – Sayeri Lala, Undergraduate Student, Massachusetts Institute of Technology
Thank You