Introduction

Proposed Method

Experiments

Detection of Cross-Channel Anomalies from Multiple Data Channels

D-S. Pham, S. Budhaditya, D. Phung, and S. Venkatesh
Curtin University
Perth, Western Australia

ICDM 2011
The Main Headlines

- "Headed for War in Iran"
- "Preparations for National Security"
- "6 Signs of Military Stance"
- "Declaring War on Iran"
- "Bush Administration Steps Up Secret Movements Against Iran"
The Minor Headlines
The Big Picture

- Information overlapping between data channels
- Topic modeling typically looking at main stories
- Cross-channel common anomalies (CCCAs)
  - Uncommon within each channel
  - Common between channels
- Research questions:
  - Can we detect these CCCAs?
  - How can we do it?
  - Theoretical analysis?
Definitions

- **Channel anomaly (CA):** anomalous document whose content differs significantly from other documents in the corpus of that channel data.

- **Cross-channel common anomaly (CCCA):** is an anomalous document with respect to a channel, but common with other anomalous documents in other channels.

Note: **CCCA must be a CA, but the converse is not true.**
Proposed Framework

Two-Stage Framework:

X_1 \rightarrow R_1 \rightarrow Z_1

X_2 \rightarrow R_2 \rightarrow Z_2

\ldots

X_N \rightarrow R_N \rightarrow Z_N

\rightarrow Z

\rightarrow C

\rightarrow CCCA Detection

\rightarrow Topic Modelling

Individual residual subspaces

Individual set of CA documents

Principal Subspace of Residuals

\begin{align*}
X_1 & \rightarrow R_1 \\
X_2 & \rightarrow R_2 \\
\vdots & \phantom{\rightarrow} \\
X_N & \rightarrow R_N \\
\end{align*}

\begin{align*}
Z & \rightarrow C \\
\end{align*}

ICDM 2011
Stage 1: Single-Channel Anomaly Detection

Table: Single-channel Anomaly Detection Algorithm

Require $X_i$, $\beta$.

Covariance Matrix Computation: $\Sigma_i = X_iX_i^T / N_i$

Eigen decomposition: $\Sigma_i = U_i \Lambda_i U_i^T$.

Extract principal eigenvectors $U_i^P$.

Compute projection to residual subspace: $P_i^R = (I - U_i^P U_i^P)^T$.

Compute the threshold $Q_i\beta$.

Project a test document onto residual subspace: $z_i[n] = P_i^R x_i[n]$.

Compute $\text{SPE}_i = \|z_i[n]\|_2^2$.

Test if $\text{SPE}_i > Q_i\beta$. 
Stage 2: Cross-Channel Anomaly Detection

<table>
<thead>
<tr>
<th>Table: Cross-channel Anomaly Detection Algorithm</th>
</tr>
</thead>
</table>

Compute $Z_i$ from Algorithm 1.
Collect CA(s) across the channel as: $Z = [Z_1, Z_2 \ldots Z_M]$ 
Compute covariance matrix: $\Sigma_Z = ZZ^T / N_Z$.
Eigen decomposition: $\Sigma_Z = U_Z \Lambda_Z U_Z^T$.
Extract Principal Eigen vectors $U_c$.
Compute projection: $P_c = U_c U_c^T$
Compute SPE statistic: $\text{SPE}_i' = \|P_c z_i[n]\|_2^2$.
Test if $\text{SPE}_i'$ statistic is larger than a suitable threshold.
Algorithm Explained

Introduction
Proposed Method
Experiments
Algorithm
Analysis

Residual Subspace
Principal Subspace

$P_P x_i$

$x_i$

$U_{Cf_i}$

$U_{Ci}$

$U_{C}$

Common cross channel topics subspace

ICDM 2011
Analysis Overview

- **Goal:** to show that combining information from multiple channels is likely to reduce false alarms
- **Proof method:** indirect proof via a study of a newly defined impurity index
- **Impurity index:** the amount of channel-specific anomalies in the residual vectors input to the second stage
- **Reduction of impurity index implies better false alarm (due to channel-specific anomalies)**
- **Quantify the reduction in terms of the number of channels** $M$
Decomposition
Results

- Impurity Index Definition
  \[ \rho = \frac{\| g + q \|_2}{\| z \|_2}. \]

- Assumptions
  - White Gaussian noise
  - Channel-specific anomalies: Bernoulli random process

Theorem

*The reduction of the impurity index of the multi-channel case over the single-channel case is given by*

\[ \rho^{SC} - \rho^{MC} = \mathcal{O}(1 - 1/\sqrt{M}) \text{ with an overwhelming probability.} \]
Compared methods

- **AC**: single-channel method with aggregated data (base-line 1)
- **SC-AVG**: average of single-channel methods (base-line 2)
- **OC-SVM**: one class SVM (classic outlier detection)
- **PCP**: principal component pursuit (sparse+low-rank modeling)
- **MC**: proposed multi-channel method
Data Collection

- 30 news feeds over 2 months
- Sources: BBC, Reuters, CNN
- External sources for controlled CCCA injections: Minti (www.minti.com.au) *breastfeeding, autism, schooling*
- Evaluation metrics
  - Receiver operating characteristics (ROC) curve
  - F-score
Detection Results: ROC Performance

Figure: Detection performance from 6 news channels
## Discovered Topics

**Table:** CCCA/CA Topics Discovered from CNN

<table>
<thead>
<tr>
<th>Setting</th>
<th>CCCA Topics</th>
<th>CA Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Injections</td>
<td>breastfeeding baby women milk feed breastfed company bra maternal look wear size fit theater baby breast formula feed want bottle milk infant health hospital breast feeding food breast birth baby hospital food mother code doctor automakers posted weak us</td>
<td>top grave album chart sting gear kloot band festive perform chard hadron love help music service show told apple stun user baby breast formula feed want bottle milk film director movie audience star machete robert chimmel die phoneix hospital car accident</td>
</tr>
<tr>
<td>Without Injections</td>
<td>automakers posted weak US hewlett packard bidding war buy storage emerging market bonds interest story invest tropical storm hermine strengthened white House stress second economic stimulus package volcano erupt Indonesia'</td>
<td>obama administration regulatory restriction korea strength military china palestinian west bank dozens hamas activists tropical storm hermine strengthened white House stress second economic stimulus package volcano erupt Indonesia' international cricket charge pakistan Salman Butt</td>
</tr>
</tbody>
</table>
Detection Results: F-score Performance

ICDM 2011
Handling Evolving Data Streams

![Graph showing AUC for different channels (MC vs SS-AVG) for News Data Experiments and Video Data Experiments.](image-url)
Dependence on Num. Channels

Recall: Theory shows impurity index reduction as $\mathcal{O}(1 - 1/\sqrt{M})$
Robustness to Weak and Uneven CCCAs

![Graph showing AUC for different channels (M) with MC]
Sensitivity to CCCA Leakage

![Bar chart showing AUC for different injected documents (r)]
Video Data Experiments Overview

- Each day of the week: a data channel
- Detect rare but repeated events
- Data collection
  - A month recording of 4 CIF video
  - Busy street scene
  - First half month for training, second half month for testing
  - Separate day-night models
  - Groundtruth verified by security
Figure: Cross-channel events in video data: bus stopping (left), loitering (middle), and car in side lane (right)
Table: Detection performance from video data: Day model (top) and Night model (bottom).

<table>
<thead>
<tr>
<th>Method</th>
<th>True Positive</th>
<th>False Positive</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>83</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>OC-SVM</td>
<td>66</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>SC-AVG</td>
<td>68</td>
<td>16</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>True Positive</th>
<th>False Positive</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>64</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>OC-SVM</td>
<td>48</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>SC-AVG</td>
<td>52</td>
<td>11</td>
<td>14</td>
</tr>
</tbody>
</table>
Conclusion

- CCCA Detection: new problem of practical interest
- Proposed a two-stage spectral algorithm for detection
- Theoretical analysis
- Experiment results consistent with theory, discover meaningful patterns
- Extension to video data with suitable modeling