Multi-task Multi-modal Models for Collective Anomaly Detection

Tsuyoshi Ide ("Ide-san"), Dzung T. Phan, J. Kalagnanam PhD, Senior Technical Staff Member
IBM Thomas J. Watson Research Center

This slides are available at ide-research.net.

IEEE International Conference on Data Mining (ICDM 2017)
Outline

- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results
Wish to build a collective monitoring solution

- You have many similar but not identical industrial assets
- You want to build an anomaly detection model for each of the assets
- Straightforward solutions have serious limitations
  - 1. Treat the systems separately. Create each model individually
    - Suffers from lack of fault examples
  - 2. Build one universal model by disregarding individuality
    - Model fit is not good
Practical requirements: Need to capture both commonality and individuality

- Capture both individuality and commonality
- Automatically capture multiple operational states
  - Real-world is not single-peaked (single-modal)
- Be robust to noise
- Be highly interpretable for diagnosis purposes
Formalizing the problem as multi-task density estimation for anomaly detection

Data

\[ \{ x_1^{(n)} \in \mathbb{R}^M \} \]

Prob. density

\[ p^1(x^1 | D) \]

all data

Anomaly score

\[ a^1(x^1) \]

System 1 (in New Orleans)

\[ \vdots \]

System s

\[ \{ x_s^{(n)} \in \mathbb{R}^M \} \]

\[ p^s(x^s | D) \]

\[ a^s(x^s) \]

\- overall
\- variable-wise

System S (in New York)

\[ \vdots \]

\[ \{ x_S^{(n)} \in \mathbb{R}^M \} \]

\[ p^S(x^S | D) \]

\[ a^S(x^S) \]
Outline

- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results
Use Gaussian graphical model (GGM)-based anomaly detection approach as the basic building block.

\[ a(x) = \begin{cases} -\ln p(x \mid D) & \text{Overall score} \\ -\ln p(x_i \mid x_{-i}, D) & \text{Variable-wise score} \end{cases} \]

\[ \max_{\Lambda} \{ \ln \det \Lambda - \text{tr}(\Sigma \Lambda) - \rho \| \Lambda \|_1 \} \]

Training data:

Sparse graphical model:

Sample covariance:

Multi-variate data:

[Ide+ SDM09] [Ide+ ICDM16]
Basic modeling strategy: Combine common pattern dictionary with individual weights

Individual sparse weights

Common dictionary of sparse graphs

System 1 (in New Orleans)

... System s

System S (in New York)

Sparse GGM 1
Sparse GGM 2
Sparse GGM K

Monitoring model for System 1

Monitoring model for System 2

Monitoring model for System S

GGM=Gaussian Graphical Model
Basic modeling strategy: Resulting model will be a sparse mixture of sparse GGM

System $s$

Gaussian mixture

\[ \sum_{k=1}^{K} \pi_k^s \mathcal{N}(x^s \mid \mu_k^s, \Lambda_k^{-1}) \]

Sparse mixture weights

(\text{= automatic determination of the number of patterns})

Sparse Gaussian graphical model

GGM=Gaussian Graphical Model
Outline

- Problem setting
- Modeling strategy
  - Model inference approach
- Experimental results
Employing a Bayesian model for multi-modal MTL

- Observation model (for the $s$-th task)
  - Gaussian mixture with task-dependent weight
    \[
    \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}^s | \mu^k, (\Lambda^k)^{-1}) z^s_k
    \]

- Sparsity enforcing priors (non-conjugate)
  - Laplace prior for the precision matrix
  - Bernoulli prior for the mixture weights
    \[
    p(\Lambda^k) = \left(\frac{\rho}{4}\right)^{M^2} \exp \left(-\frac{\rho}{2} \|\Lambda^k\|_1\right)
    \]
    \[
    p(\pi^s) = p_0^{\|\pi^s\|_0} (1 - p_0)^{G^s - \|\pi^s\|_0}
    \]

- Conjugate prior on $\{\mu^k\}$ and $\{z^s\}$
Maximizing log likelihood using variational Bayes combined with point-estimation

- **Log likelihood**

\[
L = \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{k=1}^{K} \ln \mathcal{N}(x^{s(n)} \mid \mu^k, z^{s(n)}) + \sum_{k=1}^{K} \text{Lap}(\Lambda^k \mid \rho)p(\mu^k \mid \Lambda^k) + \sum_{s=1}^{S} z^{s(n)} \ln \pi^s_k + \sum_{s=1}^{S} \ln p(\pi^s)
\]

  - Likelihood by the obs. model
  - Prior distributions

- **Use VB for** \( \{\mu^k\}, \{z^{s(n)}\} \)

- **Use point-estimate for** \( \{\Lambda^k\}, \{\pi^s\} \)
  - Results in two convex optimization problems
Maximizing log likelihood using variational Bayes combined with point-estimation

- Update sample weights
- Update cluster weights
- Update precision matrices
- Update other parameters

Use new semi-closed form solution

$$\max_{\pi^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s - \tau \| \pi^s \|_0 \right\}$$

The ratio of samples assigned to the $k$-th cluster

Solved by graphical lasso [Friedman 08]

$$\max_{\Lambda^k} \left\{ \ln |\Lambda^k| - \text{Tr}(\Lambda^k Q^k) - \frac{\rho}{N_k} \| \Lambda^k \|_1 \right\}$$

total # of samples assigned to the $k$-th cluster
Solving the L0-regularized optimization problem for mixture weights

- What is the problem of the conventional VB approach?
  - Simply differentiate w.r.t. $\pi_k^s$
  - Claims to get a sparse solution [Corduneanu+ 01]
  - But mathematically $\pi_k^s$ cannot be zero due to logarithm

- We re-formalized the problem as a convex mixed-integer programming
  - A semi-closed form solution can be derived (→ see paper)

$$\max_{\pi^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s \right\}$$

s.t. $\|\pi^s\|_1 = 1$

$$\max_{\pi^s, y^s} \sum_{k=1}^{K} \left\{ c_k^s \ln \pi_k^s - \tau y_k^s \right\} \quad \text{s.t.} \quad \sum_{k=1}^{K} \pi_k^s = 1,$$

$$y_k^s \geq \pi_k^s - \epsilon, \quad y_k^s \in \{0, 1\} \quad \text{for} \quad k = 1, \ldots, K,$$
## Comparison with possible alternatives

<table>
<thead>
<tr>
<th></th>
<th>Interpretability</th>
<th>Noise reduction</th>
<th>Fleet-readiness</th>
<th>Multi-modality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our work [Ide et al. ICDM 17]</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>(single) sparse GGM</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Gaussian mixtures</td>
<td>Limited</td>
<td>Limited</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Multi-task sparse GGM</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Multi-task learning anomaly detection</td>
<td>No</td>
<td>(depends)</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

- **Interpretability**: How well can the model be interpreted and understood?
- **Noise reduction**: Ability to reduce noise in the data.
- **Fleet-readiness**: Whether the model is suitable for large-scale applications.
- **Multi-modality**: Whether the model can handle multiple types of data inputs.
Outline

- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results
Experiment (1): Learning sparse mixture weights

- Conventional ARD approach sometimes get stuck with local minima
  - ARD = automatic relevance determination
  - Often less sparser than the proposed convex L0 approach

- Typical result of log likelihood vs VB iteration count →

Proposed convex L0 approach gives better likelihood

Conventional ARD approach gets stuck with a local minimum
Experiment (2): Learning GGMs and detecting anomalies

- “Anuran Calls” (frog voice) data in UCI Archive
  - Multi-modal (multi-peaked)
  - Voice signal + attributes (species, etc.)
- Created 10-variate, 3-task dataset
  - Use the species of “Rhinellagranulosa” as the anomaly
- Results
  - Two non-empty GGMs are automatically detected starting from $K=9$
  - Clearly outperformed single-modal MTL alternative in anomaly detection
    - Group graphical lasso, fused graphical lasso

Example of variable-wise distribution

Automatically learned GGMs
Conclusion

- Developed multi-task density estimation framework that can handle multi-modality
  - Featuring double sparsity: mixture weights, variable dependency

- Demonstrated the utility in the context of condition-based asset management
Thank you!
Integrated monitoring tool allows sharing rare anomaly data across different assets

- In condition-based monitoring, big data may not be really big
  - Anomalous samples account for less than 0.2% in a metal smelting process

- Coverage of anomalies and thus accuracy can be limited due to lack of data
Existing methods cannot handle multi-modality

Comparing the proposed multi-task multi-modal (MTL-MM) model with standard Gaussian mixture (GMM) and multi-task learning (MTL) models