Recent advances in sensor data analytics

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

- General challenges in Cognitive Manufacturing
- Change detection using directional statistics
  - T. Ide et al., IJCAI 16
- Multi-task multi-modal models for collective anomaly detection
  - T. Ide et al., ICDM 17
- Summary and future work
Cognitive Manufacturing is IBM’s research initiative to address Industry 4.0

First
Industrial machines driven by steam power

First mechanical loom, 1784

Second
Introduction of mass production systems

First conveyor belt, Cincinnati slaughterhouse, 1870

Third
Automation with electronic devices

First programmable logic controller (PLC) Modicon 084, 1969

Fourth
AI revolution: incorporation of learning system

TODAY

Degree of complexity

Fourth
Third
Second
First
Key technical areas of Cognitive Manufacturing: Sensor data analytics plays a key role

- **Anomaly / change detection**: Detect indications of failures before happening
- **Failure risk analysis**: Compute the risk of failure based on past failure records
- **Maintenance scheduling / planning**: Optimize maintenance actions while satisfying business constraints
- **Operational condition optimization**: Adjust suboptimal operational conditions
General challenges: No “one-size-fits-all” algorithm

- Example in anomaly detection
  - “Happy families are all alike; every unhappy family is unhappy in its own way.” - Anna Karenina, Leo Tolstoy

Examples of anomalies

- Outliers (from i.i.d. samples)
- Outliers (from auto-correlated samples)
- Change points
- Discords

General challenges: Business requirements often drive extension of existing approaches

- Example: corporate-level asset management with anomaly detection
  - Typically assets are managed as a cohort
    - 10s of off-shore oil production systems
    - 100s of industrial robots
    - 1000s of electric vehicles in a certain area
  - How can we leverage the commonality between assets to build an anomaly detection solution for individual assets?

General challenges: Complex internal structure may exist in one measurement

Example from semiconductor manufacturing (etching)

General challenges: Ready-to-use solution to your problem might not even exist

- Example: Charge retention (~ battery life) prediction of electric vehicle batteries
  - Depends on the entire history of battery usage
  - Battery usage is represented as a complex trajectory of a multi-dimensional space

- Charge retention prediction task should be formulated as “trajectory regression”

\[
y = f(\text{“trajectory”})
\]

General challenges: Ground truth may not be available. Some degrees of freedom are usually latent

- Example: sensor data of a compressor of oil production system
  - Data taken under a normal operational condition
  - Noisy, nonstationary, heterogeneous, high-dimensional …
- Hard to pinpoint what is indicative of malfunction
Agenda

- General challenges in Cognitive Manufacturing

- Change detection using directional statistics
  - T. Ide et al., IJCAI 16

- Multi-task multi-modal models for collective anomaly detection
  - T. Ide et al., ICDM 17

- Summary and future work
Continuous operation of conveyor systems is critical in the mining industry

- **Business goal**: Ensure continuous operation of conveyor system (“apron feeder”) by detecting early indications of failures

- **Data**: Physical sensor data from conveyors and motors
  - Every several seconds over ~ 1 year
  - Sensors include: Gearbox temperatures, motor power consumptions, apron speed, etc.

- **Challenge**: Conveyor system is subject to significant fluctuation in load. Hard to characterize the normal operation
  - Mined crude ore never come in a uniform size
Problem setting: change detection from multivariate noisy time-series data

- Change = difference between $p(x)$ and $p_t(x)$
  - $x$: $M$-dimensional i.i.d. observation
  - $p(x)$: p.d.f. estimated from training window
  - $p_t(x)$: p.d.f. estimated from the test window at time $t$

- Assume a sequence of i.i.d. vectors
  - Training data in the training window
    $\{x^{(1)}, \ldots, x^{(t)}, \ldots, x^{(N)}\}$
Problem setting: change detection from multi-variate noisy time-series data

- Question 1: What kind of model should we use for the probability density?
- Question 2: How can we quantify the difference between the densities?
We use von Mises-Fisher distribution to model $p(x)$ and $p_t(x)$

- vMF distribution: “Gaussian for unit vectors”
  \[ p(z \mid u, \kappa) = c_M(\kappa) \exp(\kappa u^\top z) \]
  - $z$: random unit vector of $\|z\| = 1$
  - $u$: mean direction
  - $\kappa$: “concentration” (~ precision in Gaussian)
  - $M$: dimensionality

- We are concerned only with the direction of observation $x$:
  - $z = \frac{x}{\|x\|}$
    - Normalization is always made
    - Do not care about the norm
    - same direction = same input
Normalization is useful to suppress multiplicative noise

- Real mechanical systems often incur **multiplicative** noise
  - Example: two belt conveyors operated by the same motor

- Normalization of vector is simple but powerful method for noise reduction
Mean direction $u$ is learned via maximum likelihood. Introduce sample weight to down-weight contaminated ones

- Weighted likelihood function

$$L(u, \kappa) = \sum_{n=1}^{N} w(n) b(n) \left\{ \ln c_M(\kappa) + \kappa u^\top z(n) \right\}$$

- Regularization over sample weights

$$R(w) = \frac{1}{2} \| w \|_2^2 + \nu \| w \|_1$$

- Parameters are learned by solving

$$(u^*, w^*) = \arg \max_{u, w} \{ L(u, \kappa) + \lambda R(w) \}$$
Multiple patterns (directions) can be obtained by coupling maximum likelihood equations

- Find orthogonal sequence of the mean direction $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ by coupling the weighted regularized maximum likelihood

\[
\begin{align*}
(u_1^*, w_1^*) &= \underset{u_1, w_1}{\text{arg max}} \{ L(u_1, \kappa) + \lambda R(w_1) \} \\
(u_2^*, w_2^*) &= \underset{u_2, w_2}{\text{arg max}} \{ L(u_2, \kappa) + \lambda R(w_2) \} \\
&\vdots \\
(u_m^*, w_m^*) &= \underset{u_m, w_m}{\text{arg max}} \{ L(u_m, \kappa) + \lambda R(w_m) \}
\end{align*}
\]

Orthogonality condition

\[
\mathbf{u}_i^\top \mathbf{u}_j = \delta_{i,j}
\]

Kronecker delta
Iterative sequential algorithm for the coupled maximum likelihood

- For each $i$, $w_i$ and $u_i$ are solved iteratively until convergence
- Analytic solution exists in each step
- Results in very simple fixed point equations

\[ w_1 \quad \cdots \quad u_1 \]
\[ w_2 \quad \cdots \quad u_2 \downarrow \quad u_1 \]
\[ \vdots \]
\[ w_m \quad \cdots \quad u_m \downarrow \quad u_i \]

$(i = 1, \ldots, m - 1)$
Derived fixed-point iteration algorithm

Example: \( i = 1 \)

Given \( w_1 \), solve

\[
\max_{u_1} \{ \kappa u_1^T X w_1 \} \quad \text{s.t.} \quad u_1^T u_1 = 1
\]

Given \( u_1 \), solve

\[
\min_{w_1} \left\{ \frac{1}{2} \| w_1 - \frac{q}{\lambda} \|_2^2 + \nu \| w_1 \|_1 \right\}
\]

\[ q \equiv \ln c_M b + \kappa X^T u_1 \]

This Lasso problem is solved analytically

---

**Algorithm 1** RED algorithm.

**Input:** Initialized \( w \). Regularization parameters \( \lambda, \nu \). Concentration parameter \( \kappa \). The number of major directional patterns \( m \).

**Output:** \( U = [u_1, \ldots, u_m] \) and \( W = [w_1, \ldots, w_m] \).

**for** \( j = 1, 2, \ldots, m \) **do**

**while** no convergence **do**

\[
u_j \leftarrow \kappa [I_M - U_{j-1} U_{j-1}^T] X w_j
\]

\[
u_j \leftarrow \text{sign}(u_j^T X w_j) \frac{u_j}{\|u_j\|_2}
\]

\[ q_j \leftarrow \gamma b + \kappa X^T u_j
\]

\[
w_j \leftarrow \text{sign}(q_j) \odot \max \left\{ \frac{|q_j|}{\lambda} - \nu 1, 0 \right\}
\]

**end while**

**end for**

Return \( U \) and \( W \).
Theoretical property: The algorithm is reduced to the “trust-region subproblem” in $\nu \rightarrow 0$

**Theorem 2.** When $\nu$ tends to 0, the nonconvex problem (5) is reduced to an optimization problem in the form of

$$\min_{u} \{u^T Qu + c^T u\} \quad s.t. \quad u^T u = 1,$$

(23)

which has a global solution obtained in polynomial time.

**Proof.** The non-convex optimization problem (23) is known as the trust region subproblem. For polynomial algorithms to the global solution, see [Sorensen, 1997; Tao and An, 1998; Hager, 2001; Toint et al., 2009]. Here we show how the algorithm is reduced to the trust region subproblem.

Useful to initialize the iterative algorithm
With extracted directions, define the change score at time $t$ as

$$a^{(t)} = \min_{f,g} \int d\mathbf{x} \mathcal{M}(\mathbf{x}|\mathbf{U}f, \kappa) \ln \frac{\mathcal{M}(\mathbf{x}|\mathbf{U}f, \kappa)}{\mathcal{M}(\mathbf{x}|\mathbf{U}^{(t)}g, \kappa)}$$

$$f^\top f = 1, \ g^\top g = 1$$

Concisely represented by the top singular value of $\mathbf{U}^\top \mathbf{U}^{(t)}$

$\mathbf{U} \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_m]$  $\mathbf{U}^{(t)} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_r]$
Experiment: Failure detection of ore belt conveyors

- vMF formulation successfully suppressed very noisy non-Gaussian noise of multiplicative nature

- ~40% of samples were automatically excluded from the model

- Better than alternatives
  - PCA, Hoteling $T^2$
  - Stationary subspace analysis [Blythe et al., 2012]
General challenges in Cognitive Manufacturing

Change detection using directional statistics
  - T. Ide et al., IJCAI 16

Multi-task multi-modal models for collective anomaly detection
  - T. Ide et al., ICDM 17

Summary and future work
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

\(\{\mathbf{x}^1(n) \in \mathbb{R}^M\}\)

\(\{\mathbf{x}^s(n) \in \mathbb{R}^M\}\)

\(\{\mathbf{x}^S(n) \in \mathbb{R}^M\}\)

Prob. density

\(p^1(\mathbf{x}^1 | \mathcal{D})\)

all data

\(p^s(\mathbf{x}^s | \mathcal{D})\)

\(p^S(\mathbf{x}^S | \mathcal{D})\)

Anomaly score

\(a^1(\mathbf{x}^1)\)

\(a^s(\mathbf{x}^s)\)

- overall
- variable-wise

\(a^S(\mathbf{x}^S)\)
Use Gaussian graphical model (GGM)-based anomaly detection approach as the basic building block

- Multi-variate data
- Sparse graphical model
- Anomaly score

\[ a(x) = \begin{cases} -\ln p(x \mid D) \\ - \ln p(x_i \mid x_{-i}, D) \end{cases} \]

Overall score
Variable-wise score

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

sample covariance

training data

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

System 1 (in New Orleans)

System s

System S (in New York)

Individual sparse weights

Common dictionary of sparse graphs

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

\[
\text{Monitoring model for System } s = \sum_{k=1}^{K} \pi_k^s \mathcal{N}(x^s | \mu_k^s, (\Lambda_k^s)^{-1})
\]

Sparse mixture weights

(= automatic determination of the number of patterns)

Sparse Gaussian graphical model

\[
\text{GGM=Gaussian Graphical Model}
\]
Propose a Bayesian multi-task model with two sparsity-enforcing priors

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

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

- **Conjugate prior on** \(\{\mu^k\}\) and \(\{z^s\}\)
\[
\begin{align*}
  p(\mu^k | \Lambda^k) &= \mathcal{N}(\mu^k | \mathbf{m^0}, (\lambda_0 \Lambda^k)^{-1}) \\
  p(z^s | \pi^s) &= \prod_{k=1}^{K} (\pi_k^s)^{z_k^s}
\end{align*}
\]
Maximizing log likelihood using variational Bayes combined with point-estimation

- Complete log likelihood

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

  - 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

\[
\max_{\pi^s_k} \left\{ \sum_{k=1}^{K} C^s_k \ln \pi^s_k - \tau \| \pi^s \|_0 \right\}
\]

The ratio of samples assigned to the \( k \)-th cluster

s.t. \( \| \pi^s \|_1 = 1 \).

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 $L_0$-regularized optimization problem for mixture weights

- Conventional VB approach without $L_0$ regularization on $\pi_k^s$ is problematic
  - Claimed 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>Method</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>

[Varoquaux et al., 2010; Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014; Gao et al., 2016; Peterson et al., 2015].

[Varoquaux et al., 2010; Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014; Gao et al., 2016; Peterson et al., 2015].
Experiment (1): Learning sparse mixture weights

- Conventional ARD approach sometimes gets stuck with local minima
  - $\text{ARD} = \text{automatic relevance determination}$
  - Often less sparse than the proposed convex $L_0$ approach

- Typical result of log likelihood vs VB iteration count

\[
\text{log-likelihood} \quad \begin{cases} 
-10^{-4} & 0 \leq \text{number of iterations} \leq 200 \\
-10^{-5} & 200 < \text{number of iterations} \\
\end{cases}
\]

- Proposed convex $L_0$ 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
Agenda

- General challenges in Cognitive Manufacturing

- Change detection using directional statistics
  - T. Ide et al., IJCAI 16

- Multi-task multi-modal models for collective anomaly detection
  - T. Ide et al., ICDM 17

- Summary and future work
Industrial sensor data have many interesting features that call for new machine learning formulation.

Introduced two recent works on anomaly detection:
- Feature extraction method based on von Mises-Fisher distribution
- Bayesian multi-task density estimation with double sparsity

Ongoing/future work
Discussion: What is the potential of deep learning in cognitive manufacturing?
Thank you!
(For ref.) Algorithm for sparse structure learning

- Assume graphical Gaussian model

\[ p(x|\Lambda) = \mathcal{N}(x|0, \Lambda^{-1}) = \frac{\det(\Lambda)^{1/2}}{(2\pi)^{M/2}} \exp \left( -\frac{1}{2} x^\top \Lambda x \right) \]

- Put a Laplace prior on Lambda

\[ p(\Lambda) = \prod_{i,j=1}^{M} \frac{\rho}{2} \exp \left( -\rho |\Lambda_{i,j}| \right) \]

\[ \text{_rho: constant controlling the strength of prior} \]

- MAP (Maximum a posteriori) estimation for Lambda

\[ \Lambda^* = \arg \max_{\Lambda} \left\{ \ln p(\Lambda) \prod_{n=1}^{N} p(x^{(n)}|\Lambda) \right\} \]

\[ = \arg \max_{\Lambda} \left\{ \ln \det \Lambda - \text{tr}(S\Lambda) - \rho ||\Lambda||_1 \right\} \]

\[ \text{S: sample covariance matrix} \]
For ref.) Anomaly scoring algorithm (for outlier analysis)

- Define the outlier score for the $i$-th variable as

$$
\text{score}_i(\mathbf{x}|\Lambda) \equiv - \ln p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_M, \Lambda)
$$

- Lambda represents a sparse structure
- $p$ is p.d.f. defined by the graphical Gaussian model

- Final result: Anomaly score of the $i$-th variable
  - Only variables connected to the $i$-th variable play a role

$$
\text{score}_i(\mathbf{x}|\Lambda) = \frac{1}{2} \ln \frac{2\pi}{\Lambda_{i,i}} + \frac{1}{2\Lambda_{i,i}} \left( \sum_{j=1}^{M} \Lambda_{i,j} x_j \right)^2
$$
Title: Recent advances in sensor data analytics

- Abstract:
- Sensor data analytics is one of the major application fields of data mining and machine learning. Typically taking real-valued time-series data from physical sensors as the input, its problem setting includes a variety of tasks depending on the application domain, not limited to the traditional regression and classification.
- This talk will first introduce technical challenges in industrial sensor data analytics. Then it will cover recent developments in machine learning algorithms in sensor data analytics. Major topics include change detection using directional statistics and multi-task extension of graph-based anomaly detection.