Expectation–Maximization Approach to Fault Diagnosis With Missing Data

Agenda

1. Introduction
2. EM Algorithm
3. Conclusion

Introduction

This paper introduces a data-driven approach for fault diagnosis in the presence of incomplete monitor data. The expectation–maximization (EM) algorithm is applied to handle missing data in order to obtain a maximum-likelihood solution for the discrete (or categorical) distribution. Because of the nature of categorical distributions,the maximization step of the EM algorithm is shown in this paper to have an easily calculated analytical solution,making this method computationally simple. An experimental study on a ball-and-tube system is investigated to demonstrate advantages of the proposed approach.

Introduction

FDI methods

- model-based approaches

  Principle :calculate the residuals through observers,use the global analytical redundancy relations.

  Disadvantage:the applicability of these methods is limited, as building a model for more complex systems is often considered difficult.

- data-driven approaches

  Make use of historical data for training

  Bayesian solution:deal with complete training data

  EM:deal with incomplete data and multiple missing data patterns

Introduction

- Components and Behavioral Modes
  - m1=sensors.m2=actuators.m3=pipes
  - the number of modes will exponentially grow with respect to the number of components
- Evidence
  - $e \in \{e1, e2, \ldots, eK\}$
  - The number “1” means there is an abnormal event detected by the corresponding monitor; “0” indicates no abnormality has been detected.
- Training Data
  - $D = \{d1, d2, \ldots, dN\}$, D:historical training data
  - $D = \{Dm1, Dm2, \ldots, DmQ\}$

Introduction

$$D_c = \{d_c^1, d_c^2, \ldots, d_c^{N_c}\} \quad (8)$$

$$D_{\mathrm{ic}} = \left\{d_{\mathrm{ic}}^1, d_{\mathrm{ic}}^2, \ldots, d_{\mathrm{ic}}^{N_{\mathrm{ic}}}\right\} \quad (9)$$

where N_c is the total number of complete data entries corresponding to mode m_j, and N_{ic} is the number of incomplete data entries. x and z are composed of observed and missing monitor readings in D_{ic}; x and z are structured as follows:

$$\mathrm{x} = \{x_1, x_2, \ldots, x_{N_{\mathrm{ic}}}\} \quad \mathrm{z} = \{z_1, z_2, \ldots, z_{N_{\mathrm{ic}}}\}.$$

For the sake of applying the EM algorithm, the likelihood of the training data D should be expressed in terms of x and z. The likelihood of D can be written as

$$p(D|\Theta) = p(D_c, D_{\mathrm{ic}}|\Theta) = p(D_c|\Theta)p(D_{\mathrm{ic}}|\Theta). \quad (10)$$

Based on the i.i.d. assumption and (5)

$$p(D|\Theta) = \prod_{k=1}^{K} \theta_k^{n(e_k|D_c)} \prod_{t=1}^{N_{\mathrm{ic}}} p\left(d_{\mathrm{ic}}^t|\Theta\right). \quad (11)$$

The term $p(d_{\mathrm{ic}}^t|\Theta)$ can now be replaced by the joint probability $p(z_t, x_t|\Theta)$, which yields the following result:

$$p(D|\Theta) = \prod_{k=1}^{K} \theta_k^{n(e_k|D_c)} \prod_{t=1}^{N_{\mathrm{ic}}} p(z_t, x_t|\Theta) \quad (12)$$

with its log likelihood being

$$L(D|\Theta) = \log p(D_c, D_{\mathrm{ic}}|\Theta) = \sum_{k=1}^{K} n(e_k|D_c)\log\theta_k + \sum_{t=1}^{N_{\mathrm{ic}}} \log p(z_t, x_t|\Theta). \quad (13)$$

EM Algorithm

· An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables

· The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters

· A maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step

EM Algorithm

- Start, $p = 0$: set initial value of Θ^0
- Iterate (until convergence)

  The E-step: Calculate

$$Q(\Theta|\Theta^p) = E_{C_{\text{mis}}|C_{\text{obs}},\Theta^p}\left[\log\left(p(C_{\text{mis}}, C_{\text{obs}}|\Theta)\right)\right] \tag{14}$$

  where C_{obs} is the observed data set, and C_{mis} is the missing values or unobserved latent data.

  The M-step: solve

$$\Theta^{p+1} = \arg\max_{\Theta} Q(\Theta|\Theta^p). \tag{15}$$

Graphical Models with unobserved variables

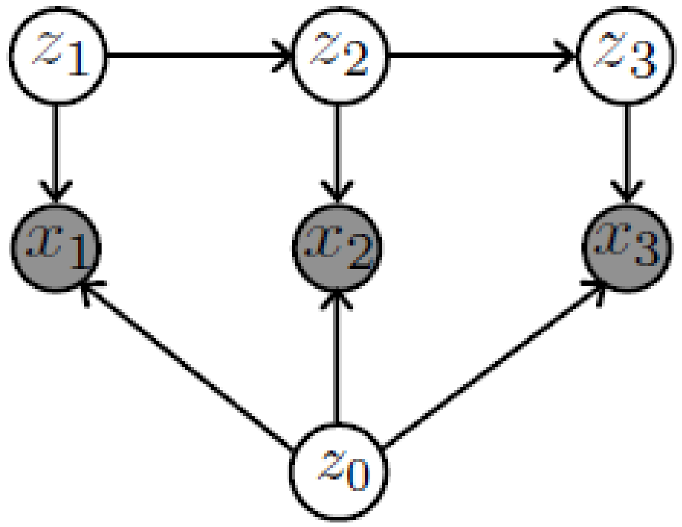
- For a directed graphical model:

θ → *fixes conditional distributions of every child node, given parents*

x → *observed nodes (training data)*

z → *unobserved nodes (hidden data)*

Inference: Find summary statistics of posterior needed for following M-step



EM Algorithm

$$\sum_{j=1}^{+\infty} P(X = x_i, Y = y_i) = \mathrm{P}(\mathrm{X} = x_i)$$

$$L(\theta) = \sum_i \ln p\left(x^{(i)};\theta\right) = \sum_i In \sum_{z^{(i)}} P(x^{(i)}, z^{(i)};\theta)$$

$$= \sum_i \ln \sum_{z^{(i)}} Q_i(z^i) \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^i)} = \sum_i \ln(E\left[\frac{P\left(x^{(i)}, z^{(i)};\theta\right)}{Q_i(z^i)}\right])$$

$$\geq \sum_i E\left[ln \frac{P\left(x^{(i)}, z^{(i)};\theta\right)}{Q_i(z^i)}\right] \geq \sum_i \sum_{z^i} Q_i\left(z^i\right) ln \frac{P\left(x^{(i)}, z^{(i)};\theta\right)}{Q_i(z^i)}$$

EM Algorithm

等号成立的条件是当且仅当$\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^i)}$=C

$$\frac{P\left(x^{(i)},z^{(i)};\theta\right)}{Q_i(z^i)}=C$$

Q_i表示隐含变量Z的某种分布，
Q_i满足的条件是：
$\sum_Z Q_i\left(z^i\right)$=1　$Q_i\left(z^i\right)\geq 0$

$$\Rightarrow\quad P\left(x^{(i)},z^{(i)};\theta\right)=\mathrm{C}(Q_i\left(z^i\right))$$

$$\Rightarrow\quad \sum_Z P\left(x^{(i)},z^{(i)};\theta\right)=C(\sum_Z Q_i\left(z^i\right))\quad\Rightarrow\quad \sum_Z P\left(x^{(i)},z^{(i)};\theta\right)=C$$

EM Algorithm

等号成立的条件是当且仅当$\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^i)}$=C

$$Q_i(z^i) = \frac{P(x^{(i)}, z^{(i)}; \theta)}{\sum_Z P(x^{(i)}, z^{(i)}; \theta)}$$

$$= \frac{P(x^{(i)}, z^{(i)}; \theta)}{P(x^{(i)}; \theta)}$$

$$= P(z^{(i)} | x^{(i)}; \theta)$$

EM Algorithm

$$Q(\theta|\theta^p) = \mathrm{L}(\theta) = \sum_i \sum_{z^i} Q_i(z^i) ln \frac{P\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i(z^i)}$$

$$= E_{C_{mis}|C_{obs},\theta^p}[\log(p(C_{mis}, C_{obs}|\theta))]$$

所以$Q(\theta|\theta^p) = E_{C_{mis}|C_{obs},\theta^p}[\log(p(C_{mis}, C_{obs}|\theta))]$

EM Algorithm

$$Q(\Theta|\Theta^p) = \sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{x}, D_c, \Theta^p) \log p(\mathbf{x}, \mathbf{z}, D_c|\Theta) \quad (16)$$

~~where $\mathbf{Z}$ is the sample space for all the possible values of the realization $\mathbf{z}$.~~ Summation over $\mathbf{Z}$ ($\sum_{\mathbf{Z}}$) implies the summation over all the possible realizations $\mathbf{z}$ in $\mathbf{Z}$. Θ is the parameter set given in (3) with the constraint $\sum \Theta = 1$. The term $p(\mathbf{z}|\mathbf{x}, D_c, \Theta^p)$ can be expressed as

$$p(\mathbf{z}|\mathbf{x}, D_c, \Theta^p) = \frac{p(\mathbf{z}, \mathbf{x}|D_c, \Theta^p)}{p(\mathbf{x}|D_c, \Theta^p)} = \frac{p(\mathbf{z}, \mathbf{x}|D_c, \Theta^p)}{\sum_{\mathbf{Z}} p(\mathbf{z}, \mathbf{x}|D_c, \Theta^p)}.$$

Because of independence between D_c and D_{ic}, (16) reduces to

$$Q(\Theta|\Theta^p) = \sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{x}, \Theta^p) \left[\log p(\mathbf{x}, \mathbf{z}|\Theta) + \log p(D_c|\Theta)\right]. \quad (17)$$

Furthermore, due to independence between D_c and $\mathbf{z}$, and the fact that $\sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{x}, \Theta^p) = 1$, the Q-function can be also expressed as

$$Q(\Theta|\Theta^p) = \log p(D_c|\Theta) + \sum_{\mathbf{Z}} p(\mathbf{z}|\mathbf{x}, \Theta^p) \log p(\mathbf{x}, \mathbf{z}|\Theta). \quad (18)$$

$$Q(\Theta|\Theta^p) = \log p(D_c|\Theta) + \sum_{Z_1}\sum_{Z_2}\cdots\sum_{Z_{N_{ic}}} \left[\prod_{t=1}^{N_{ic}} p(z_t|x_t, \Theta^p)\right]\left[\sum_{i=1}^{N_{ic}} \log p(x_i, z_i|\Theta)\right] \quad (19)$$

where $Z_1, Z_2, \ldots, Z_{N_{ic}}$ indicate the sample space for $z_1, z_2, \ldots, z_{N_{ic}}$. Now, because each $p(z_t|x_t, \Theta^p)$ sums to 1 and is constant with respect to $z_{i \neq t}$

$$\sum_{Z_1}\sum_{Z_2}\cdots\sum_{Z_{N_{ic}}} \left[\prod_{t=1}^{N_{ic}} p(z_t|x_t, \Theta^p)\right] \log p(x_{i=t}, z_{i=t}|\Theta) = \sum_{Z_t} p(z_t|x_t, \Theta^p) \log p(x_t, z_t|\Theta).$$

Thus,

$$\begin{aligned} Q(\Theta|\Theta^p) &= \log p(D_c|\Theta) + \sum_{Z_1} \left[p(z_1|x_1, \Theta^p) \log p(z_1|\Theta)\right] \\ &\quad + \cdots + \sum_{Z_{N_{ic}}} \left[p\left(z_{N_{ic}}|x_{N_{ic}}, \Theta^p\right) \log p\left(z_{N_{ic}}|\Theta\right)\right] \\ &= \log p(D_c|\Theta) + \sum_{t=1}^{N_{ic}}\sum_{Z_t} \left[p(z_t|x_t, \Theta^p) \log p(z_t, x_t|\Theta)\right]. \end{aligned} \quad (20)$$

EM Algorithm

$$Q(\theta_k|\Theta^p) = \sum_{k=1}^{K}\sum_{t=1}^{N_c} p\left(e_k|d_c^t\right)\cdot\log\theta_k + \sum_{k=1}^{K}\sum_{t=1}^{N_{ic}} p\left(e_k|d_{ic}^t,\Theta^p\right)\cdot\log\theta_k$$
$$= \sum_{k=1}^{K}\left[\sum_{t=1}^{N_c} p\left(e_k|d_c^t\right)\cdot\log\theta_k + \sum_{t=1}^{N_{ic}} p\left(e_k|d_{ic}^t,\Theta^p\right)\cdot\log\theta_k\right] \quad (21)$$

where

$$p\left(e_k|d_c^t\right) = \begin{cases} 1, & e_k = d_c^t \\ 0, & e_k \neq d_c^t \end{cases} \quad (22)$$

$$p\left(e_k|d_{ic}^t,\Theta^p\right) = \begin{cases} \frac{p(e_k|\theta^p)}{\sum_{e_j\in d_{ic}^t} p(e_j|\theta^p)}, & e_k \in d_{ic}^t \\ 0, & e_k \notin d_{ic}^t. \end{cases} \quad (23)$$

For compactness, we can write the expression in vector form and denote $\mathcal{E} = [e_1, e_2, \ldots, e_K]^T$, so that the structure with respect to Θ is clearly seen, i.e.,

$$Q(\Theta|\Theta^p) = \left[\sum_{t=1}^{N_c} p\left(\mathcal{E}|d_c^t\right) + \sum_{t=1}^{N_{ic}} p\left(\mathcal{E}|d_{ic}^t,\Theta^p\right)\right]^T \log\Theta \quad (24)$$

where $\Theta = [\theta_1, \theta_2, \ldots, \theta_K]^T$ and

$$p\left(\mathcal{E}|d_c^t\right) = \begin{bmatrix} p\left(e_1|d_c^t\right) \\ p\left(e_2|d_c^t\right) \\ \vdots \\ p\left(e_K|d_c^t\right) \end{bmatrix} \quad p\left(\mathcal{E}|d_{ic}^t,\Theta^p\right) = \begin{bmatrix} p\left(e_1|d_{ic}^t,\Theta^p\right) \\ p\left(e_2|d_{ic}^t,\Theta^p\right) \\ \vdots \\ p\left(e_K|d_{ic}^t,\Theta^p\right) \end{bmatrix}.$$

For further simplification, one can define the sum of probabilities over t based on the complete data set D_c as

$$n(e_k|D_c) = \sum_{t=1}^{N_c} p\left(e_k|d_c^t\right). \quad (25)$$

Considering in (22), $p(e_k|d_c^t)$ can only take value 0 or 1; it is the same as counting the number of complete samples in D_c, which has been explained in Section II-D. The sum of probabilities based on the incomplete data d_{ic} is regarded as the expected realization frequency for e_k, which can be denoted by

$$n(e_k|D_{ic}) = \sum_{t=1}^{N_{ic}} p\left(e_k|d_{ic}^t\right). \quad (26)$$

By applying this to the vector $p(\mathcal{E}|d_{ic}^t,\Theta^p)$

$$\sum_{t=1}^{N_{ic}} p\left(\mathcal{E}|d_{ic}^t,\Theta^p\right) = n(\mathcal{E}|D_{ic},\Theta^p). \quad (27)$$

EM Algorithm

The Q-function is expressed as

$$Q(\Theta|\Theta^p) = [n(\mathcal{E}|D_c) + n(\mathcal{E}|D_{\text{ic}}, \Theta^p)]^T \log \Theta. \quad (28)$$

From (28), it is clear that Θ can be individually optimized over each element θ_k in Θ. However, one has to keep in mind that the elements of Θ are probabilities that must sum to 1, and each element θ_k has $0 \leq \theta_k \leq 1$. Thus, the constraint $\theta_k = 1 - \bar{\theta}_k$ must be applied, where $\bar{\theta}_k = \sum_{j \neq k} \theta_j$, i.e.,

$$\begin{aligned} Q(\theta_k|\Theta^p) = {} & [n(e_k|D_c) + n(e_k|D_{\text{ic}}, \Theta^p)] \cdot \log \theta_k \\ & + [n(\bar{e}_k|D_c) + n(\bar{e}_k|D_{\text{ic}}, \Theta^p)] \cdot \log(1 - \theta_k) \quad (29) \end{aligned}$$

where the terms $n(\bar{e}_k|D_c)$ and $n(\bar{e}_k|D_{\text{ic}}, \Theta^p)$ have likewise definition, i.e.,

$$n(\bar{e}_k|D_c) = \sum_{j \neq k} n(e_j|D_c), n(\bar{e}_k|D_{\text{ic}}, \Theta^p) = \sum_{j \neq k} n(e_j|D_{\text{ic}}, \Theta^p).$$

By taking the first derivative and setting it to zero, we have

$$\begin{aligned} \frac{dQ(\theta_k|\Theta^p)}{d\theta_k} = {} & \frac{n(e_k|D_c) + n(e_k|D_{\text{ic}}, \Theta^p)}{\theta_k} \\ & - \frac{n(\bar{e}_k|D_c) + n(\bar{e}_k|D_{\text{ic}}, \Theta^p)}{1 - \theta_k} = 0. \quad (30) \end{aligned}$$

$$\begin{aligned} \theta_k^{p+1} & = \frac{n(e_k|D_c) + n(e_k|D_{\text{ic}}, \Theta^p)}{n(e_k|D_c) + n(e_k|D_{\text{ic}}, \Theta^p) + n(\bar{e}_k|D_c) + n(\bar{e}_k|D_{\text{ic}}, \Theta^p)} \\ & = \frac{n(e_k|D_c) + n(e_k|D_{\text{ic}}, \Theta^p)}{N_c + N_{\text{ic}}}. \quad (31) \end{aligned}$$

EM Algorithm

Advantages:

1) Missing data do not need to be ignored, and including this information increases accuracy of diagnosis.

2) The proposed method is able to handle the multiple missing data patterns.

3) the EM algorithm converges quickly. Particularly in the single missing data pattern problem, it takes one iteration to converge to the result.

EM Algorithm

Limitations:

1) The EMalgorithm only guarantees local convergence, not necessarily a globally optimal solution.

2) The EM algorithm for missing data will suffer from an overfitting problem in higher dimensions.

3) This solution assumes that the cause of missing data is independent of the mode. If certain modes cause some particular missing pattern or lead more data to be missing,then this method may not work or at least no longer optimal.

Thanks!