Advanced in Control Engineering and Information Science

The Mutual Information based Correlation Analysis between Fault types and Monitor Data
Huo Lin\textsuperscript{a}, Lv Chuan\textsuperscript{a}, Zhou Dong\textsuperscript{a}, Fei Simiao\textsuperscript{b}, a*

\textsuperscript{a}State Key Laboratory of Virtual Reality Technology and System, BeiHang University, Beijing100191, China
\textsuperscript{b}Dept. of system Engineering of Engineering Technology, BeiHang University, Beijing100191, China

Abstract

The fault monitor data analysis of aviation equipment is essential for the failure diagnosis and further researches. Here we propose a new approach based on Mutual Information to measure the correlation of the fault types and the data indexes, particularly in how to analysis the correlation of discrete variables and continuous variables. Then the fire control system of aircraft is taken for example to make sure the fault symptom class. Furthermore, we compare the results with that by Correlation Coefficient method and draw the conclusion that the Mutual Information method based on entropy makes more sense.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of [CEIS 2011]

Keywords: correlation analysis, mutual information, correlation coefficient, entropy

1. Introduction

The analysis, disposal and selection of different kinds of monitor data are crucial to further researches in the complex system or the aviation equipment. The correlation study of the monitor data and the system or equipment fault can provide basis for the failure diagnosis to understand the operation condition of equipment more exactly and efficiently. Therefore, how to analyse the correlation quantitative between fault types and monitor data is very important and most researches are still in the qualitative phase so far.

Many familiar methods such as correlation analysis and regression analysis are broadly used in the correlation analysis study whereas have their own limits [1, 2]. Although the studies on the correlation analysis domestic and overseas go deepen continually, the theory and method of entropy has its own advantages—entropy is the uncertainty measurement of the random variable. The method of Mutual Information based on entropy could measure all the statistical correlation between statistical variables and has been applied in the medicine field [3]. Some researchers combined Mutual Information and

* Corresponding author. Tel.: +1-521-098-2335
E-mail address: helen0404@yahoo.cn.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2011.08.976
Correlation Coefficient to describe the correlate degree. Conant R.C. used the entropy theory to divide the complex system [4]. Chen Jing discussed the application of entropy and its evolution in data correlation analysis [5]. Tonoi and Edelman successfully zoned the brain function by entropy in the medicine [6]. Xi Guangcheng proposed the least two rules for the ideal dividing requirement—the correlation of each divided subsystem should have Limited Additiviy and Countable Additiviy, and analysed the correlation of the ecological economy region and divided into zones [7, 8]. Sun Z.Q. studied the correlation between TCM syndromes and physicochemical parameters [9].

This paper first introduces the usual Correlation Coefficient method in brief and then focuses the Mutual Information based on entropy to measure the correlation between discrete variables and continuous variables. The Mutual Information method is limited in the correlation analysis between discrete variables in majority so far. Here we study the correlation between fault types and monitor data of complex system or equipment and take the fire control system of aircraft for example to make the fault symptom class and then compare the results with that by the Correlation Coefficient method, concluding that the Mutual Information method is more reasonable.

2. The Correlative Coefficient Method

Many Correlation Analysis methods including Correlation Coefficient method, the Logistic Regression Analysis method and so on are widely used in every walk of life. Here in this paper we introduce the Correlation Coefficient method in brief. The correlation coefficient is the index to measure the statistical correlation of at least two random variables.

In the correlation study of random variable $X$ and $Y$, we assume Means as $E(X)=\mu$, $E(Y)=\nu$ and variations as $\sigma^2_X, \sigma^2_Y$. The standard-variables of $X$ and $Y$ are $U = \frac{X - \mu}{\sigma_X}$, $V = \frac{Y - \nu}{\sigma_Y}$ and $E(U) = 0, E(V) = 0, D(U) = 1, D(V) = 1$. $\rho_{XY}$ is the correlation coefficient of $X$ and $Y$ which represents the Mean of multiplying the standard variable $X$ and $Y$.

$$\rho_{XY} = E(UV) = E\left[\left(\frac{X - \mu}{\sigma_X}\right)\left(\frac{Y - \nu}{\sigma_Y}\right)\right]$$

In the Eq.1, $-1 \leq \rho_{XY} \leq 1$. $\rho_{XY}$ is a plus quantity which means $X$ and $Y$ are positively correlate, and vice versa. The Correlation Coefficient method limited in the application as it could only reveal the linear correlation between variables.

3. The Entropy Based Mutual Information Method

3.1. Mutual Information Based on Entropy

Entropy is the disordered degree of the system and has broadly applied in Cybernetics, Probability, Cosmical Physics and Bionomy. Entropy reveals the uncertainty of object in the information system.

Mutual information is a useful information measurement and is based on entropy, which shows the correlation of two incident class. The mutual information between $X$ and $Y$ is defined as the entropy of the measure $\mu_{XY}$ of $\mu_X \times \mu_Y$, and is showed as $I(X,Y) = H(\mu_{XY}; \mu_X \times \mu_Y)$. Many definition format of entropy has been proposed and the most widely used entropy is Shannon Entropy. [10, 11]
Assumed $X_1, X_2, \cdots, X_n$ as characteristic variables and its probability density functions are $p(x^1), p(x^2), \cdots, p(x^n)$. The definition domain of $X_i (i=1,2,\cdots,n)$ is $\Omega_i$ and the joint probability densities are described as $p(x^i, x^j) (i=1,2,\cdots,n; j=1,2,\cdots,n)$.

When $X_i (i=1,2,\cdots,n)$ is continuous variable, the Shannon entropy of $X_i$ is
\[
H(X_i) = -\int_{\Omega_i} p(x^i) \ln p(x^i) dx^i 
\]
When $X_i (i=1,2,\cdots,n)$ is discrete variable and $X_j (j=1,2,\cdots,n)$ is continuous variable, the joint entropy of $X_i$ and $X_j$ is
\[
H(X_i, X_j) = -\sum_{x^i \in \Omega_i} \int_{\Omega_j} p(x^i, x^j) \ln p(x^i, x^j) dx^j 
\]
The mutual information between $X_i$ and $X_j$ can be expressed as
\[
I(X_i; X_j) = H(X_i) + H(X_j) - H(X_i, X_j) 
\]

3.2. Correlation Study between Discrete Variables and Continuous Variables

When considering the correlation between fault types and data indexes, the fault types of aircraft can be seen as discrete random variables, while monitoring data can be seen as continuous random variables. Mutual Information method is usually limited to the studies of the correlation of discrete random variables instead of continuous random variables. Here we propose a new approach to quality the correlation between fault type and monitoring data.

After random sample is given, it is particular important to find appropriate way to conform the distribution and estimate the parameters. If Conditional Probability Density Function (CPDF) has known, question would be much simpler. Here we assume that the CPDF of monitoring data is Normal Distribution as we know from history experience. The parameters of Normal Distribution can be estimated by maximum likelihood method.

Given a set of continuous characteristic variables $X_j$ and discrete variables $C \in (c_1, c_2, \cdots, c_j)$, we define $x^j_i \in R (j=1,2,\cdots,N)$, $c_i \in \{1,2,\cdots,J\}$ and get sample $(x^1_i, x^2_i, \cdots, x^n_i, c_i)$, $i=1,2,\cdots,J, x_i \in R^n$ and aim to quantify the correlation between characteristic random variables and type variables. The CPDF of continuous random variables is $p(x^i | C = c_j) (i=1,2,\cdots,n; j=1,2,\cdots,J)$, which we assume as Normal Distribution. When the parameters of CPDF have been given, we can classify the sample into $J$ groups. There are $N_j$ elements each subsample and $\sum_{j=1}^{J} N_j = N$. The parameters of Normal Distribution $\mu_j, \sigma_j (i=1,2,\cdots,n; j=1,2,\cdots,J)$ can be estimated by using maximum likelihood method.

The CPDF is $p(x^i | C = c_j) = \frac{1}{\sqrt{2\pi} \sigma_j} \exp\left(-\frac{(x^i - \mu_j)^2}{2\sigma_j^2}\right)$.

The PDF (Probability Density Function) of discrete variables is $P(C = c_j) = \frac{N_j}{N}$, $j=1,2,\cdots,J$.

The joint density function of $X_i$ and $C = j$ is
The marginal density function of $X_i$ is

$$p(x^i) = \sum_{j=1}^{N} p(x^i | C = j) = \frac{N_j}{\sqrt{2\pi N \sigma_{i,j}}} \exp\left(-\frac{(x^i - \mu_{i,j})^2}{2\sigma_{i,j}^2}\right)$$

The mutual information of $X_i$ and $C = j$ can be expressed as

$$I(C, X_i) = \sum_{j=1}^{N} \int p(x^i | C = j) \ln\left(\frac{p(x^i | C = j)}{p(x^i)P(C = j)}\right) dx^i$$

$$= \sum_{j=1}^{N} \frac{N_j}{N} \ln(\sqrt{2\pi \sigma_{i,j}}) - \frac{1}{2} - H(X_i)$$

The Shannon entropy of $X_i$ can be got from Eq.2.

Mean and variance can be get from maximum likelihood method as Eq.4.

$$\mu_i = \frac{1}{N} \sum_{i=1}^{N} x_i, \quad \sigma_i^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_i)^2$$

4. Example

The fault of fire control system is one of fault type when aircraft operating. There are many detection indexes which could be used for failure diagnosis. So it would be a representative example for research of correlation between fault types and data indexes.

There are 15 key voltage detection points which could describe 11 kinds of fault type in aircraft fire control system. Table 1 shows a part of the sample of monitoring voltage data for the lack of space. The data contains 11 kind of fault including antennae fault, high frequency receiver fault, master oscillator fault, transmitter fault, synchronizer fault, control unit fault, switch unit fault, adjusting unit fault, computer fault, power unit fault.

Each components fault could be described as binary variable. And these 11 kinds of fault type are express as $S_1, S_2, \ldots, S_{11}$. When the fault type belongs to the component $i$, $S_i = 1$, otherwise $S_i = 0 (i = 1, 2, \ldots, 11)$. The data records 15 detection points’ voltage and express by $X_1, X_2, \ldots, X_{15}$. With these data we can study the correlation between fault types and monitoring data. As we do not consider the correlation between index data, so we can analyse every single correlation between voltage data each component fault. the volume of sample is confirmed by the record of $X_j$ in correlation analysis.

Table 1 The monitor fault data of aircraft fire control system

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
<th>$S_6$</th>
<th>$S_7$</th>
<th>$S_8$</th>
<th>$S_9$</th>
<th>$S_{10}$</th>
<th>$S_{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.23</td>
<td>4.80</td>
<td>4.78</td>
<td>1.33</td>
<td>4.80</td>
<td>4.79</td>
<td>4.80</td>
<td>4.90</td>
<td>4.80</td>
<td>0.03</td>
<td>4.80</td>
</tr>
<tr>
<td>0.28</td>
<td>4.78</td>
<td>4.84</td>
<td>1.21</td>
<td>4.86</td>
<td>4.71</td>
<td>4.56</td>
<td>2.12</td>
<td>4.92</td>
<td>0.10</td>
<td>4.55</td>
</tr>
<tr>
<td>0.21</td>
<td>4.44</td>
<td>4.47</td>
<td>1.44</td>
<td>4.77</td>
<td>4.68</td>
<td>4.68</td>
<td>2.45</td>
<td>4.64</td>
<td>0.06</td>
<td>4.68</td>
</tr>
<tr>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23.48</td>
<td>23.49</td>
<td>23.50</td>
<td>23.50</td>
<td>23.50</td>
<td>23.49</td>
<td>23.49</td>
<td>23.10</td>
<td>23.50</td>
<td>0.10</td>
<td>1.01</td>
</tr>
<tr>
<td>23.46</td>
<td>23.33</td>
<td>23.45</td>
<td>23.49</td>
<td>23.57</td>
<td>23.48</td>
<td>23.46</td>
<td>23.25</td>
<td>23.55</td>
<td>0.12</td>
<td>1.02</td>
</tr>
<tr>
<td>23.21</td>
<td>23.48</td>
<td>23.53</td>
<td>23.45</td>
<td>23.52</td>
<td>23.51</td>
<td>23.48</td>
<td>23.21</td>
<td>23.54</td>
<td>0.10</td>
<td>1.01</td>
</tr>
<tr>
<td>$x_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.01</td>
<td>0.02</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.06</td>
<td>0.02</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>0.06</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.05</td>
<td>0.01</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
</tr>
</tbody>
</table>
4.1. Mutual Information Method.

The sample volume of the monitor random data $X_j$ is $N_j$ ($j=1,2,\cdots,15$), and when $S_i=1$ and $S_j=0$ ($i=1,2,\cdots,12$), the sample volume are respective $N_{ij}^0$ and $N_{ij}^1$. So the PDF of fault type $S_i$ is $P_{ij}^k = N_{ij}^k / N_{ij}$ ($k=0,1$). The conditional distribution of $X_j$ is assumed as Normal Distribution and $\mu, \sigma$ could be estimated by maximum likelihood method. After calculating the mutual information using Eq.2-4, we can select the higher correlation data and get maximum correlation classes below:

\[
(X_1, S_1), (X_2, S_2), (X_3, S_3), (X_4, S_4), (X_5, S_5), (X_6, S_6), (X_7, S_7), (X_8, S_8), (X_9, S_9), (X_{10}, S_{10}), (X_{11}, S_{11}), (X_{12}, S_{12}), (X_{13}, S_{13}), (X_{14}, S_{14}), (X_{15}, S_{15})
\]

Table 2 The correlation form by correlation information method

<table>
<thead>
<tr>
<th>Fault Type</th>
<th>Maximum Correlation Class</th>
<th>Fault Type</th>
<th>Maximum Correlation Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$X_4$</td>
<td>$S_7$</td>
<td>$X_{11,15}$</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$X_4, X_{11}$</td>
<td>$S_8$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$X_4$</td>
<td>$S_9$</td>
<td>$X_6$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$X_4$</td>
<td>$S_{10}$</td>
<td>$X_{10}$</td>
</tr>
<tr>
<td>$S_5$</td>
<td>$X_4$</td>
<td>$S_{11}$</td>
<td>$X_{6,12,14}$</td>
</tr>
<tr>
<td>$S_6$</td>
<td>$X_4$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As Table2 shows, the ones which have bigger correlation with $S_{11}$ are $X_9, X_{12}, X_{14}$. In other words, when we diagnose this kind of fault, the bigger correlation can support more information. So we can say that the symptom of $S_{11}$ fault contains $X_9, X_{12}, X_{14}$.

4.2. Correlation Coefficient Method

We can get the results of correlation analysis between fault types and monitor data using Eq.1, as it is shown in Table 3.
<table>
<thead>
<tr>
<th>Fault Type</th>
<th>Maximum Correlation Class</th>
<th>Fault Type</th>
<th>Maximum Correlation Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>X4,X12,X14</td>
<td>S7</td>
<td>X11,X15</td>
</tr>
<tr>
<td>S2</td>
<td>X1,X13</td>
<td>S8</td>
<td>X1</td>
</tr>
<tr>
<td>S3</td>
<td>X5</td>
<td>S9</td>
<td>X8</td>
</tr>
<tr>
<td>S4</td>
<td>X6</td>
<td>S10</td>
<td>X10</td>
</tr>
<tr>
<td>S5</td>
<td>X2</td>
<td>S11</td>
<td>X9</td>
</tr>
<tr>
<td>S6</td>
<td>X7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Obviously, the outputs of Correlation Coefficient and Mutual Information have the similar results. The main difference occurs in fault symptom of S11. However, by analysing fault mechanism, we find out that the fault of power unit fault has impact on monitoring points X9, X12, X14 directly. Correlation Coefficient method is limited to study the relationship of liner variables instead of Mutual Information method which can quantity the correlation of complex relationship variables. Therefore, the Mutual Information method can describe the real relationship of variables better.

5. Conclusion

In summary, we analysed the correlation between fault types and monitor data of the complex system or equipment by Mutual Information based on entropy, particularly between discrete variables and continuous variables and then applied into the failure diagnosis of the aircraft fire control system. The comparison of results between Mutual Information and Correlation Coefficient reveals that the results by Mutual Information method were more practical and reasonable, therefore could provide more accurate information for the equipment failure diagnosis.

However, the entropy based Mutual Information method in this study has its own limits. Our conditional probability distribution model of continuous variables here is based on Normal Distribution and this assumption is reasonable in some conditions. After we have found the essential fault symptom data, further work will be required in how to select the optimal number and class—optimal fault symptom class.

Reference