Data Missing Bayesian Network Parameters Learning Optimization Algorithm Based on EM

Zhao-Jing TONG\textsuperscript{1,2,a,*}, Yun-Ji ZHAO\textsuperscript{1,b}, Rui-Jun TAN\textsuperscript{2,c}, Jun-Ling SHI\textsuperscript{1,d}

\textsuperscript{1}School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, 454000, China

\textsuperscript{2}Henan Pingyuan Optics & Electronics. Co. Ltd, Jiaozuo 454001, China

\textsuperscript{a}tong6661@163.com, \textsuperscript{b}auyjz@hpu.edu.cn, \textsuperscript{c}trj258@sina.com, \textsuperscript{d}897063266@qq.com

*Corresponding author

Keydwords: Bayesian Networks, Data Missing, Parameter Learning.

Abstract. The paper proposes an optimization algorithm for the parameter learning of missing data Bayesian networks. Expectation Maximization (EM) algorithm is the common parameter learning algorithms. The maximum likelihood estimation (MLE) and maximum a posterity estimation (MAP) of EM are local estimate rather than global estimate and are not easy to achieve the global optimal. So this paper puts forward a point estimate relative error minimum optimization algorithm which based on EM algorithm (EM-MLE-MAP). Applying the algorithm to the fault diagnosis of the rotor vibration Bayesian network, simulations and experiments show that the improved algorithm has good precision in diagnosing vibration fault when the loss ratio less than 3 percent.

Introduction

Bayesian network as a very useful tool in data mining can provide qualitative and quantitative relationship between attributes and probability inference. It is one of the most effective theoretical models to achieve the uncertain knowledge expression and reasoning in the field of artificial intelligence [1]. Bayesian network becomes a hot spot in the field of fault diagnosis with solid theoretical foundation, the natural form of expression, flexible reasoning ability and convenient decision-making mechanism [2]. The data missing BNs parameters learning is an important aspect of BNs research.

Data missing BNs parameter learning is a studying for the parameters which in view of the background in which the data is incomplete or cannot be observed. The poor of data is the normal problem that we have to deal with in measuring, part of which caused by inaccurate data with noise or measurement, and the other part is data missing, namely the integrity of data. For the first case, we can use some hardware and software to eliminate or improve it. For the latter data missing, the current widespread application is Gibbs sampling algorithm and Expectation Maximization algorithm. It is feasible by using the above methods to search local minimum. But it is difficult for universal search, and it assumed that the data missing mechanism is negligible. But when the data missing can’t be neglected, the precision of above algorithm have reduced significantly.

EM algorithm is a kind of estimate method of gradual certainty for the unknown parameters of data missing when assumes that missing value is independent of the observations [3]. There are two cases after got the estimation $\hat{\theta}$ of point $\theta$: one is the
maximum likelihood estimation MLE which makes the $p(e | \theta)$ reach maximum; the other is maximum posterior estimation which makes the $p(e | \theta)$ reach maximum. But it is a kind of local estimation rather than global estimation when estimated. We can’t get the global optimal while the two estimation may get local optimal solution. The paper proposed an improved algorithm based on the learning of EM-MLE and EM-MAP algorithms.

**Traditional EM Estimation Algorithms**

Although the EM algorithm can deal with data missing, it assumes that the data source from a Gaussian parameters model or mixed model, and it can only deal with data of numerical attribute. And its convergence speed is slow. The commonly used EM algorithm of data missing uses Expectation Step and Maximization Step alternatively when the BNs parameter learning to achieve local optimal of the data likelihood value.

**The Process of EM and EM-MLE Algorithms**

1. Initializing the $\theta^{(0)}$, selecting the desired accuracy $\varepsilon$ for $\theta$, requiring the revised set to the larger value $\theta$ at least one cycle time. When $|\hat{\theta} - \theta| > \varepsilon$, $\theta \rightarrow \theta$;

2. The probability distribution of missing date (E-Step)

$$P(e' | e, \theta) = \frac{P(e | e', \theta)P(e' | \theta)}{\sum_{e'} P(e | e', \theta)P(e' | \theta)}$$

3. Under the condition of $P(e' | e, \theta)$, the maximum likelihood estimation or maximum a posteriority estimation of $\theta$ (M-Step); step 2 formula (1) calculate the expected sufficient statistic $s$ of missing value $e'$ rather than itself distribution. $s$ is the sufficient statistic of $\theta$ relative to $e'$ if and only if $\theta$ and the missing date $e'$ independent of each other.

EM-MLE algorithm: ① Initializing the $\theta^{(0)}$, selecting the desired accuracy $\varepsilon$ for $\theta$, requiring the revised set to the larger value $\theta$ at least one cycle time. When $|\hat{\theta} - \theta| > \varepsilon$, $\theta \rightarrow \theta$; ② Calculate the sufficient statistic of the expected missing value

$$E_{P(X|y, \theta)} N_{ijk} = \sum_{i} P(X_{ik}, parents(X_{ij}) | y, \theta)$$

In the formula (2) : $N_{ijk}$ is all the joint instantiation count of $X_{ik}, parents(X_{ij})$, among them respectively with $k, j$ to show the count of $X_{ij}, parents(X_{ij})$. Because there are estimation parameters $\theta$ and a causal structure in this step, we can calculate Right Hand Side (RHS) probability by using the BN. We can obtain any $N_{ijk}$ through summing the Right Hand Side (RHS) probability of all the joint observation.

3. Using expected maximize statistics $P(e' | \theta)$;

$$\hat{\theta}_{ijk} = \frac{E_{P(X|y, \theta)} N_{ijk}}{\sum_{i} E_{P(X|y, \theta)} N_{ijk}}$$

112
EM-MAP Algorithm

The initialization of EM estimation algorithm based on the MAP is the same as the MLE of EM algorithm, in other words, the initialization $\theta^{(0)}$ to $\theta$ selection the expected accuracy $\epsilon$, and requiring setting the revised $\theta$. Formula (1) is to calculate the missing value $e'$ of the expected sufficient statistic. What different is its expression of calculating the maximum expected statistics in the end. The MAP is to calculate the maximum expected statistics $P(\theta | e')$ which is the same as real values.

$$\theta_{\phi} = \frac{\alpha_{i\phi} \cdot E_{p(\|e'|\theta)} N_{i\phi}}{\sum_k (\alpha_{i\phi} + E_{p(\|e'|\theta)} N_{i\phi})}$$ (4)

The Improved EM-MLE-MAP Algorithm

EM algorithm which applied to parameter learning of BNs use Expectation Step (E-Step) alternatively, according to the existing parameter find complete data of expected and the Maximum Step (M-Step), reevaluating parameters again, adjusting parameters through iterative, so as to achieve local optimization of the data likelihood value. BNs parameters learning of conventional data missing of the EM estimation algorithm in the M step, to calculate the MLE value or MAP, the difference of data missing of the physical process has a certain influence on the calculation results of the two which is accuracy of point estimation. The paper determine the algorithm selection mode through selecting and analyzing the measured sample of the two algorithms MLE and MAP to contrast calculation to estimate relative error under the condition of different data missing: Assume that no missing sample "existence" missing value, and then use MLE and MAP algorithm to estimate the missing value respectively, we select the relative calculation error between the smallest as the estimation algorithm of such samples from the two algorithm, finally we complete the real estimation of data missing. This algorithm is based on EM-MLE and MAP of the point estimation; it is called the EM-MLE-MAP algorithm. EM-MLE-MAP based on the point estimation calculate the expectation of missing value $e'$ sufficient statistic in the first through the formula (2), and then to test the actual samples relative error comparison respectively with formula (3) and formula (4) the MLE value and MAP to determine the choice of less as a practical calculation model, finally we obtain the point estimation value $\theta$ more accurate.

Improved EM-MLE-MAP Algorithms Applied in Fault Diagnosis BNs

The Establishment of the Network Fault Diagnosis with the Rotor Vibration BNS

Because of the complexity, functions and work processes of the diesel generators, there are many causes and symptoms. For its common rotor vibration faults: rotor misalignment, rubbing, unbalance, bearing looseness, etc. Its related symptoms indicate characteristics from the following generally: the main vibration frequency distribution, the change of vibration with the speed, the change of vibration with the load, etc. Consider the relationship between all-cause and all-symptom, we choose the appropriate nodes to establish preliminary Bayesian network for fault diagnosis. The model of vibration fault diagnosis contains 12 nodes which is divided into three layers. The paper built Bayesian Network structure directly with our experience and general reasoning. Domain knowledge can be written into knowledge base in
the form of rules, and we can get the BNs structure to express the issues by using programs reasoning of constraint logic. To verify the validity of the algorithm, parameter learning of data missing BNs which has been built originally, the algorithm uses the existing study structure to restudy "data missing samples", it uses the learned new network to reason fault probability, then compare their accuracy.

Contrast of Diagnosis Effect before and after Algorithm Optimized

The paper uses the EM-MLE, the EM-MAP and EM-MLE-MAP algorithm to study parameter learning with the date missing samples in chart 1 by respectively based on BNT under the environment of MATLAB and toolbox.

Figure-1 shows the minimum optimization algorithm (EM-MLE-MAP) of point estimation relative error which based on EM for the data missing of BNs parameter learning, Figure-2 shows parameter learning based on BNs, and to show the Contrast of diagnosis accuracy before and after algorithm optimized.

As Figure 1 show, compared with diagnostic accuracy given the BN parameter learning in EM-MLE, EM-MAP, and EM-MLE-MAP, the fault diagnostic accuracy of the parameter in EM-MLE-MAP was in a quite high level of about 98% and changes steadily.

In order to determine the influence of data missing rate to the accuracy of algorithm, this paper increasing the amount of data missing respectively by random, and then statistic the fault diagnosis accuracy after studying the EM-MLE-MAP algorithm parameters learning. The statistical data in the following chart.

<table>
<thead>
<tr>
<th>Rate of the missing data(%)</th>
<th>1</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (% )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM-MLE</td>
<td>93.8</td>
<td>87</td>
<td>89.2</td>
<td>91</td>
<td>85</td>
<td>86</td>
<td>75.4</td>
<td>74</td>
<td>68</td>
<td>45</td>
</tr>
<tr>
<td>EM-MAP</td>
<td>94</td>
<td>90</td>
<td>93</td>
<td>97</td>
<td>80.5</td>
<td>78</td>
<td>65</td>
<td>56</td>
<td>47</td>
<td>63</td>
</tr>
<tr>
<td>EM-MLE-MAP</td>
<td>98</td>
<td>96.9</td>
<td>96.7</td>
<td>96.5</td>
<td>94</td>
<td>90</td>
<td>78</td>
<td>72</td>
<td>60</td>
<td>50</td>
</tr>
</tbody>
</table>
From the data of table-1, we can see parameter learning algorithm of EM-MLE-MAP get relatively high diagnostic accuracy compared with EM-MLE and EM-MAP under the different rate of data missing. But with the increase of the rate of data missing, the overall accuracy decline rapidly. When the percentage of data missing reached 1%, the EM-MLE-MAP algorithm diagnosis accuracy reached 98%. If the missing rate less than 3%, the diagnosis accuracy of algorithm is more than 96.5%. As the rate of date missing rises again, fault misdiagnosis rate get increased. So we can obtain more accurate diagnosis if the rates of missing are controlled in a smaller range in the engineering application.

Conclusion

This paper presents the BN parameters learning of deficiency data based on the EM-MLE-MAP diagnostic whose relative error is the least. Uniting the parameter study of some diesel generator’s rotor turning Bayesian fault diagnostic internet, the paper experiment setting parameter learning lack of the data between 1%~8% and reasons and verifies the arithmetic of EM-MLE-MAP. The simulation experiment shows that when the rate of the data missing is less than 3%, EM-MLE-MAP can achieve a higher diagnostic accuracy rate. The diagnostic solves the problem of low-data lack rate BNs parameter learning quite well, and suits for the applying of the fault diagnostic engineering.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (U1504623); Key Laboratory of control engineering of Henan Province(KG2014-12); Key Laboratory of Mine Informatization (KY2015-07), Henan Polytechnic University.

References:


