ACSIJ Advances in Computer Science: an International Journal, Vol. 2, Issue 5, No.6 , November 2013
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Automatic identification of corrosive factors categories according to the environmental factors

Qing Xu[1], Dongmei Fu[2]

**[1] School of Automation and Electrical Engineering,
University of Science and Technology Beijing,
Beijing 100083, China**
xuqing8917@163.com

**[2] School of Automation and Electrical Engineering,
University of Science and Technology Beijing,
Beijing 100083, China**
fdm2003@163.com

Abstract

Time of wetness, SO_2 and Cl^- pollutants are three key factors for the selection of metal materials in engineering applications and the determination of atmospheric corrosivity categories. In the past, when one or more corrosive factors data is missing, corrosive factors categories were often subjectively determined according to expert experience. In order to overcome such difficulty, this paper presents a method to automatically determine corrosive factors categories using detected environmental factors data instead of expert scoring. In this method, Bayesian network was used to build the mathematical model. And the inference was obtained by clique tree algorithm. The validity of the model and algorithm was verified by the simulation results.

Keywords: *Bayesian Network, Clique tree algorithm, Corrosive factors, Environmental factors.*

1. Introduction

The increase in the numbers of types of metal materials has brought more options for engineering applications. However, the associated differences in material performances also produced difficulties in the selection and application of these materials. Due to the expanding applications of the metal materials and the complexity in application environment, accurate evaluation of atmospheric corrosivity categories is of significant importance to the research in the corrosion behaviors of metal materials and their selections [1]. At present, classification of atmospheric corrosivity of metals is completed according to ISO 9223:1992(E) [2]. In this standard, the pollution categories and time of wetness categories are first classified based on the actual measured values of the key data including time of wetness, concentration of sulfur dioxide and concentration of chloride. Then atmospheric corrosivity categories can be determined by synthesizing pollution categories and time of wetness categories together. However, using ISO 9223:9223 (E) can be very difficult due to the lack of accuracy in long-term monitoring of the numerical values of time of wetness, concentration of sulfur dioxide and concentration of chloride and the limited availability of the area containing all three data.

Environmental factors including mean annual temperature (^oC), annual average rainfall (ml), annual average sunshine (h/a), wind speed (m/s) and location (i.e. whether it is inshore) are in close correlation with corrosive factors (i.e. time of wetness, concentration of sulfur dioxide and concentration of chloride) but can be more easily measured. Thus, they were often used in the past for expert scoring when the corrosive factors data were missing. The expert determination process of corrosive factors categories is shown in Fig. 1.

To overcome the limitations of subjective expert scoring, this study proposed a computer automatic method to determine corrosive factors categories by using Bayesian network [3] and clique tree algorithm [4] on the basis of the five aforementioned environmental factors (Fig. 2). From Fig. 1 and Fig. 2, the main content of this study included: (1) finding the relationship of corrosive factors and environmental factors, and then structuring the Bayesian network model; (2) auto-differentiating corrosive factors in the data missing situation.

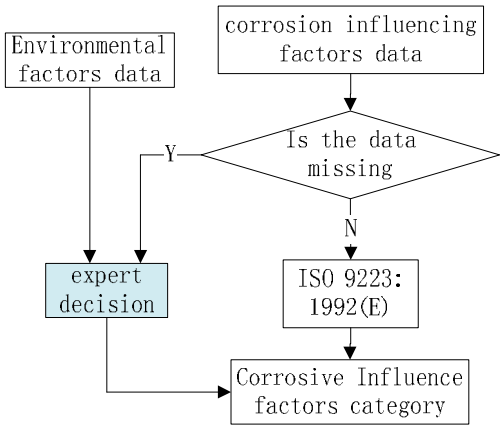

Fig. 1 Experts scoring to determine corrosive factors categories.

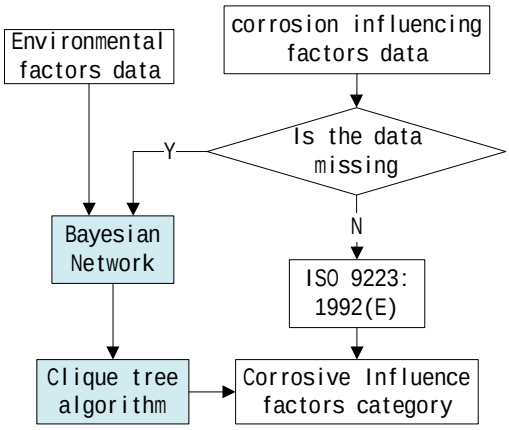

Fig. 2 Computer automatic determining corrosive factors categories.

Bayesian network [5] is a graphical model to describe dependencies between random variables, consisting of directed acyclic graph and conditional probability tables. Directed acyclic graph represents qualitative dependent relationship between variables. And conditional probability tables (root node corresponding marginal distribution) represent quantitative relationship between variables.

Clique tree is an undirected tree. Every node in the tree is called group node, which is the most connected sub graph of undirected graph and is composed of a set of random variables in the Bayesian network.

In this study, 78 corrosive environment data samples from different areas of the world were used to build the Bayesian network structure models of the three corrosive factors and five environmental factors. These samples were representatively selected from China meteorological data sharing service system and the national environmental corrosion materials science data sharing service system to cover 78 cities around the world and different climatic types including typical corrosive climates.

The computer automatic method was comprised of four parts: 1) processing the data sample into standard data set suitable for building Bayesian network model of the three corrosive factors and five environmental factors; 2) obtaining Bayesian network model according to the standard data; 3) transforming Bayesian network structure into clique tree; 4) based on the measured values of environmental factors, using the clique tree reference to obtain corrosive factors categories. The technical route of this study is shown in Fig. 3.

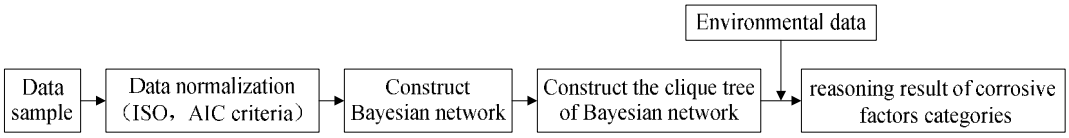

Fig. 3 The flow chart of technical route.

2. Data normalization

Constructing Bayesian network with continuous variables can be very difficult. Currently, the most commonly used approach is converting continuous variables into discrete variables [6] and then constructing Bayesian network model. Data discretization can be completed by either directly discretizing continuous variables or clustering the sampling data. In this study, the discretization process is described as below:

According to ISO 9223:1992(E), SO_2 pollution, Cl^- pollution and time of wetness for standard outdoor atmospheres were categorized by discretizing the measured values of concentration of sulfur dioxide, concentration of chlorides in the atmospheres, and time of wetness on the specimen, respectively (Table 1-3).

Table 1: SO_2 pollution categories [2]

SO_2 category	Deposition rate of SO_2 $mg/m^2 \cdot d$	Concentration of SO_2 mg/m^3	Example of occurrence
P_0	$P_d < 10$	$P_c \le 12$	The city of Wanning
P_1	$10 < P_d \le 35$	$12 < P_c \le 40$	The city of Shenzhen
P_2	$35 < P_d \le 80$	$40 < P_c \le 90$	The city of Wuhan
P_3	$80 < P_d \le 200$	$90 < P_c \le 250$	The city of Moscow

Table 2: Cl^- pollution categories [2]

Cl^- category	Concentration of Cl^- $mg/m^2 \cdot d$	Example of occurrence
S_0	$S < 3$	The city of Beijing
S_1	$3 < S \le 60$	The city of Qionghai
S_2	$60 < S \le 300$	The city of Qingdao
S_3	$300 < S \le 1500$	The city of Singapore

The four environmental factors (i.e. mean annual temperature, annual average rainfall, annual average sunshine and wind speed) can be clustered and classified using nearest neighbor clustering algorithm based on AIC criterion [7].

$$AIC = 2\sum_{m=1}^{N} \frac{d_{\max} - d_{\min}}{Q(m)} + 2N\left[1 + \ln\frac{K}{N}\right] \quad (K > N) \tag{1}$$

$$d_{\max} = \max\{D_m \mid m = 1,2,\mathbf{L}, N\} \tag{2}$$

$$d_{\min} = \min\{D_m \mid m = 1,2,\mathbf{L}, N\} \tag{3}$$

in which, K is the number of sample data, N is the number of categories, $Q(m)$ is the number of m-class samples, D_m is the within-class deviation of m-class samples, m=1，2，…，N.

Table 3: time of wetness categories [2]

category	time of wetness		Example of occurrence
	h/a	%	
t_1	$t \le 10$	$t \le 0.1$	The city of Aswan
t_2	$10 < t \le 250$	$0.1 < t \le 3$	The city of Beijing
t_3	$250 < t \le 2500$	$3 < t \le 30$	The city of Tokyo
t_4	$2500 < t \le 5500$	$30 < t \le 60$	The city of Guangzhou
t_5	$t > 5500$	$t > 60$	The city of Wanning

From the 78 corrosive environment data samples mentioned earlier, the optimal classification numbers and the corresponding AIC values of all the factors are given in Table 4.

Integrating Table 1-4, discrete levels of corrosion environment variables are summarized in Table 5.

Table 4: Optimal clustering results of the variables

cluster variables	mean annual temperature	annual average rainfall	annual average sunshine	annual average sunshine
clustering number	5	6	3	6
AIC value	6.5221	15.6476	15.5103	2.8762

Table 5: Number of corrosion environment variable state

environment variables	time of wetness categories	SO_2 pollution categories	Cl^- pollution categories	mean annual temperature	annual average rainfall	annual average sunshine	wind speed	inshore
clustering number	5	4	4	5	6	3	6	2

3. Constructing Bayesian network model of corrosive factors and environmental factors

Constructing Bayesian network structure with small data set mainly included four steps: 1) building maximum likelihood tree based on small data set; 2) extending small data set, combining the maximum likelihood tree with Gibbs sampling to revise the extended data set, and then obtaining the complete data set; 3) orienting the maximum likelihood tree based on small samples to obtain the directed tree, then using the directed tree and complete directed acyclic graph to sort block nodes and nodes in

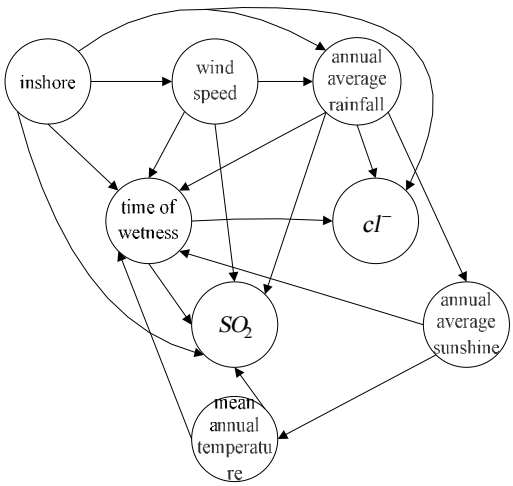

Fig. 4 Bayesian network structure.

different blocks; 4) based on sorted nodes, using the local scale - search method[8] to construct Bayesian network structure of the complete data set.

In Fig. 4, the Bayesian network structure based on the 78 data samples in this study is constructed. The nodes represent corrosion environment variables and the directed edges represent the relationship between any two variables. This network structure qualitatively describes the close relationship between corrosive factors and environmental factors.

4. Automatic identification of the corrosive factors categories based on clique tree algorithm

Bayesian network structure qualitatively expresses the conditional independent relationships between the nodes in the network. When father nodes of one node are given, the rest of the nodes are conditional independent with that node except for its subsequent nodes. Based on the conditional independent relationship between the nodes, joint probability distribution expressed by Bayesian network can be simplified in the following form [9]:

$$P(X_1, X_2, \mathbf{L}, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \quad (4)$$

in which, $Pa(X_i)$ is the father node set of the node X_i.

In Bayesian network inference, the conditional probability $P(X \mid E)$ is calculated based on conditional independent relationships in the network. Here, X is the interrogation node, and E is the given evidence node. To date, clique tree algorithm is the most popular precise inference algorithm among all Bayesian network inference algorithms [6]. In clique tree algorithm joint probability distribution is expressed by graphical presentation of the clique tree. Its basic idea is to convert Bayesian network into clique tree, followed by evidential reasoning via the messaging process defined on the clique tree.

4.1 Constructing the clique tree of Bayesian network

The classic Bayesian network inference algorithms [10] - probability propagation in trees of clusters (also known as clique tree algorithm) was established in 1988 by Lauritzen and Spiegelhalter and improved in 1992 by Dawid [11-12]. Based on clique tree algorithms, the Bayesian network structure (Fig. 4) can be converted to the clique tree (Fig. 5), in which the circular nodes are called group node and rectangle nodes at the intersections of two adjacent group nodes are called segmentation nodes.

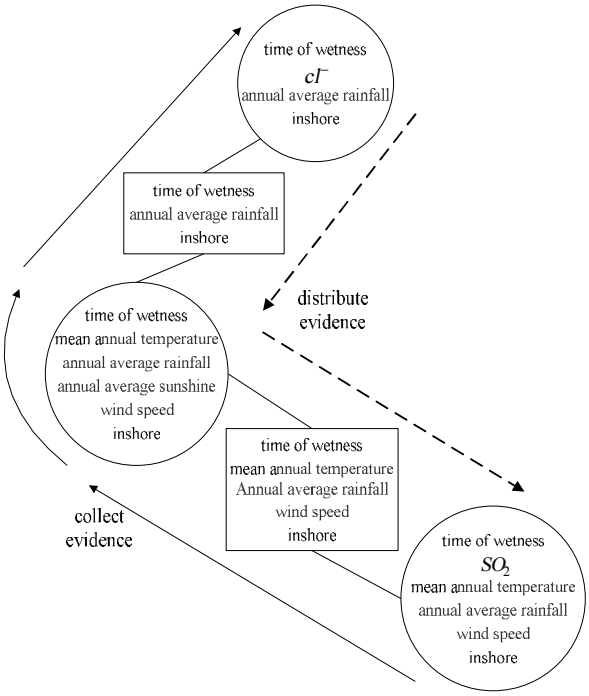

Fig. 5 Clique tree of Bayesian network shown in Fig.4.

4.2 Automatic identification of the corrosive factors categories

The process of evidential reasoning based on the clique tree is divided into two phases [6]: collecting evidence and distributing evidence. In Fig.5, time of wetness, Cl^-, annual average rainfall and location are selected as root nodes. Solid and dotted arrows indicate the processes of

collecting evidence and distributing evidence, respectively. According to the clique tree algorithm, the measured values of environmental variables are inputted to the clique tree as known evidence. The process of collecting evidence starts from the most distant node from the root node and the evidence is distributed through each node towards the root node. Upon receiving evidence, the node modifies its parameters and passes the evidence to the next one till the root node. On the other hand, the process of distributing evidence starts from the root node. The received evidence on each node is inputted to its parameters and then passed to the adjacent node till the evidence has reached every node. After that, from any clique node containing interrogation node, the probability distribution can be calculated through marginalization.

Using environmental data of Guangzhou as a test sample, we demonstrated and tested the reasoning process of the clique tree. The values of corrosion environment factors in Guangzhou are given in Table 6.

Table 6: Corrosive environment factors of Guangzhou

City	time of wetness categories	SO_2 pollution categories	Cl^- pollution categories	mean annual temperature	annual average rainfall	annual average sunshine	wind speed	inshore
Guangzhou	t_4	P_1	S_2	22	1736.1	1773.2	1.7	Yes

The last five variables in Table 6 were inputted to the clique tree as evidence variables and the first three as interrogation variables. The obtained results of reasoning are given in Table 7.

Table 7: The reasoning results of Guangzhou corrosive factors

Corrosive factors	The value of each variable and the corresponding probability				
time of wetness categories	t_1	t_2	t_3	t_4	t_5
	0	0	0.1429	**0.7143**	0.1429
SO_2 pollution categories	P_0	P_1	P_2	P_3	
	0.1345	**0.4051**	0.3133	0.1472	
Cl^- pollution categories	S_0	S_1	S_2	S_3	
	0.1238	0.1279	**0.4385**	0.3099	

In Table 7, the corrosive factors category which corresponds to the maximum probability of corrosive factors variables is the most possible category in the given environment. By using the clique tree in Fig.5, the following results were thereby produced: the most possible categories for the time of wetness (t_4), SO_2 pollution (P_1) and Cl^- pollution (S_2) were consistent with the measured values in Table 6. This method was applied onto the 78 data samples for this study and statistical analysis was performed on obtained reasoning results (Table 8).

Table 8: The reasoning coincide of 78 data records

Coincide between reasoning results and actual values	three factors are the same	two factors are the same	one factor is the same	all are not the same
Reasoning correct number	45	23	10	0
percentage	57.69%	29.49%	12.82%	0

Table 8 shows the coincidence between the reasoning results of three corrosive factors and the actual data. In the reasoning results, the probability that three corrosive factors are the same as the actual values is 57.69%; the probability that two corrosive factors are the same as the actual values is 29.49%; the probability that only one corrosive factor is same as the actual value is 12.82%; the situation where all of the corrosive factors are totally different from the actual values does not exist. Although discrepancies exist, the reasoning results are still very close to the actual value. In Table 8 and Table 9, Qingdao and Shenyang were used as examples for the comparisons between the actual values and the reasoning results.

Table 9: The reasoning results of Qingdao corrosive factors

Corrosive factors	The value of each variable and the corresponding probability					Actual value
time of wetness categories	t_1	t_2	t_3	t_4	t_5	t_4
	0	0	0	**1**	0	
SO_2 pollution categories	P_0	P_1	P_2	P_3		P_2
	0.0645	**0.3871**	0.3548	0.1935		
Cl^- pollution categories	S_0	S_1	S_2	S_3		S_2
	0.0645	0.1290	**0.4839**	0.3226		

Table 9 lists the actual corrosive factors categories and the reasoning corrosive factors categories for the example of Qingdao. Two of corrosive factors categories (i.e. time of wetness category and Cl^- pollution category) are the same as the actual values. For SO_2 pollution category, the actual value is P_2 whereas the reasoning result is P_1. Despite such difference, the reasoning result is still close to the actual value and can be used as a good reference.

Table 10: The reasoning results of Shenyang corrosive factors

Corrosive factors	The value of each variable and the corresponding probability					Actual value
time of wetness categories	t_1	t_2	t_3	t_4	t_5	t_3
	0	0.1429	**0.4434**	0.4137	0	
SO_2 pollution categories	P_0	P_1	P_2	P_3		P_2
	0.1531	**0.4278**	***0.2776***	0.1415		
Cl^- pollution categories	S_0	S_1	S_2	S_3		S_1
	0.3178	***0.1624***	0.3145	0.2052		

In the case of Shenyang (Table 10), only time of wetness category is the same as the actual value. SO_2 pollution category and Cl^- pollution category from the reasoning do not agree with the actual values, but are sufficiently close to be still considered as practical references.

The results showed that the percentage of correctly reasoning two and more corrosive factors categories reached 87.18%. Overall, when these three corrosive factors are difficult to detect or test cost is high, using the method in this study, the corrosive factors categories can be reasoned according to the detected environmental data.

5. Conclusions

To summarize, a new approach for automatic determination of corrosive factors categories was proposed using Bayesian network and clique tree algorithm. Simulation results confirmed that the reasoning accuracy of proposed approach can basically meet the actual demand when corrosive factors categories in the application environment of metal materials are determined. The proposed approach provided a scientific basis for the corrosion protection of metal materials, and established a solid foundation for automatic determination of atmospheric corrosivity categories and the corresponding materials selection.

Acknowledgments

First of all, I would like to extend my sincere gratitude to my supervisor, Fu Dong-mei, for her instructive advice and useful suggestions on my thesis. I am deeply grateful of her help in the completion of this thesis. Special thanks should go to my friends who have put considerable time and effort into their comments on the draft. Finally, I am indebted to my parents for their continuous support and encouragement.

References

[1] Yong-Ji Weng，Xiang-Yi Li．Corrision prediction and basic chemometrics[M]．Beijing：Petroleum Industry Press，2011．

[2] IS0 9223-1992(E) Corrosion of metals, and alloys - Corrosivity of atmospheres – Classification．

[3] WANG Shuang-Cheng, LENG Cui-Ping, LI Xiao-Lin．Learning Bayesian Network Structure from Small Data Set [J]．Acta Automatica Sinica，2009,35(8)：1063-1070．

[4] Wei-Na Liu, Li-Min Huo, Li-Guo Zhang．Research of exact inference algorithm in Bayesian Networks [J]．Microcomputer Information，2006,22(3-3)：92-94．

[5] Pearl J．Probablilistic Reasoning in Intelligent Systems: Networks of Plausible Inference．California: Morgan Kaufmann,1988,383-408．

[6] Shuang-Cheng Wang．Bayesian network learning, inference and application [M]．Shanghai：Lixin Accounting Press，2009．

[7] Xuan-Yun Qin．Nearest neighbor clustering algorithm based on AIC criterion [J]．Systems Engineering and Electronics，2005,27(2)：257-259．

[8] France．PSI Technical Reports．BNT Structure Learning Package: Documentation and Experiments, PB08-76801[R]．

[9] Bensi, Michelle；Kiureghian, Armen Der；Straub, Danie．Efficient Bayesian Net － work modeling of systems[J]．RELIABILITY ENGINEERING & SYSTEM SAFETY ,2013,112：200-213．

[10] LAURITZEN S L. SPIEGELHALTER D J．Local computations with probabilities on graphical structures and their application to expert systems [J]．Journal of the Royal Statistical Society, 1988, 50(2):157-224．

[11] Cowell R.G. Dawid A.P．Fast retraction of evidence in a probabilistic expert system [J]．Statistics and Computing, 1992, 2(1): 37-40．

[12] Dawid A.P．Applications of a general propagation algorithm for probabilistic expert systems[J]．Statistics and Computing, 1992, 2(1): 25-36．