Hybrid Reliability Parameter Selection Method Based on Text Mining, Frequent Pattern Growth Algorithm and Fuzzy Bayesian Network

SHUAI Yong1,2∗ (帅勇), SONG Tailiang3 (宋太亮), WANG Jianping1 (王建平), ZHAN Wenbin4 (詹文斌)
(1. Technical Support Engineering Faculty, Armored Forced Engineering Academy, Beijing 100072, China; 2. Unit 68207 of PLA, Jiayuguan 735100, Gansu, China; 3. China Defense Science & Technology Information Center, Beijing 100142, China; 4. Communication Station, Air Force Equipment Research Academy, Beijing 100085, China)

© Shanghai Jiao Tong University and Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract: Reliability parameter selection is very important in the period of equipment project design and demonstration. In this paper, the problem in selecting the reliability parameters and their number is proposed. In order to solve this problem, the thought of text mining is used to extract the feature and curtail feature sets from text data firstly, and frequent pattern tree (FPT) of the text data is constructed to reason frequent item-set between the key factors by frequent pattern growth (FPG) algorithm. Then on the basis of fuzzy Bayesian network (FBN) and sample distribution, this paper fuzzifies the key attributes, which forms associated relationship in frequent item-sets and their main parameters, eliminates the subjective influence factors and obtains condition mutual information and maximum weight directed tree among all the attribute variables. Furthermore, the hybrid model is established by reason fuzzy prior probability and contingent probability and concluding parameter learning method. Finally, the example indicates the model is believable and effective.

Key words: reliability parameter, text mining, frequent pattern growth (FPG), fuzzy Bayesian network (FBN)

CLC number: TP 311  Document code: A

0 Introduction

In the period of equipment design scheme and demonstration, the designer will determine the equipment reliability parameters according to system engineering principle and the warfare characteristics of equipment. To different types of equipment, the way to select reliability parameters and their number mainly includes the following three points. The first method is the expert scoring combined with other methods. For example, Zhen and Li[1] combined expert scoring method with Bayesian theory. Zhou and Zhu[2] proposed quality function expansion method, while this sort of method was too subjective and reliant to experts judgment. The second method is the design of experiments. For example, Lu and Jin[3] proposed the method of orthogonal experimental design, and thus this method had hysteresis quality in parameter selection and could not judge the number of parameters. The third one is using probability and frequency to determine the parameters. For example, Shao et al.[4] proposed a diversiform sampling plan; this kind of method did not generalize the parameters and could not find the relationship between the data. In view of the existing problems, this paper adopts text mining, association rules analysis and fuzzy Bayesian network (FBN) to create a hybrid parameter selection model.

1 Relevant Theories

The concept of text mining is different from data mining, it mainly focuses on extracting useful knowledge from unstructured or semi-structured text, and data mining mainly finds main mode of data from structured database[5]. The main use of text mining is to extract the unknown knowledge from the text. Because text mining must process the unstructured text data, it covers information technology, text analysis, pattern recognition, statistics, data visualization, database technology, machine learning and data mining technology, which is a complex subject area. General process of text mining is shown in Fig. 1.

Frequent pattern growth (FPG) algorithm uses a specially designed storage structure called frequent pattern
tree (FPT) to store transaction data. This tree structure compresses database storage space greatly with the store mode of common node. FPG is an algorithm that searches all frequent item-sets directly without generating candidate item-sets on frequent tree[6]. FPG algorithm uses a dividing and conquering strategy, and the experiment shows that it is faster than Apriori algorithm. Bayesian network is graphics mode used to represent connective relationship probability among the variables, and it is a main method to solve the uncertainty knowledge reasoning, which provides a natural and causal information expressing method to discover the potential relationship of data[7–10]. Ambiguity reflects the subjective uncertainty, which is the uncertainty to the definition or the concept of some problems in the language sense.

The whole modeling process is shown in Fig. 2. Combining the above idea of data mining algorithms, this paper represents and extracts the features of text data by text mining technology from text, obtains the representation represents documents by some feature terms, and it only needs to process these feature items during text mining so as to process unstructured text.

2 Parameter Selection Modeling

The key factors need to be mined, belonging to set \( \{X_1, X_2, \cdots, X_n\} \), in which \( n \) is the total number of key factors. Every parameter described by the key factors belongs to the set \( \{(f_{i1}, x_{11}), (x_{21}, x_{22}, \cdots, x_{2j}), \cdots (x_{nk})\}, n, i, j, k \in \mathbb{Z}. \)

2.1 Text Mining and Feature Set Curtailing Model

The main steps of text mining and feature set curtailing model are presented as follows.

(1) Text data preprocessing method. This paper takes the reverse maximum method to preprocess the text data[5]. The method chooses a character string which contains 6 to 8 Chinese characters as the maximum character string and matches the maximum character string with the word entries in the dictionary. If they cannot match, then cut off a Chinese character and continue matching until find corresponding word location in the dictionary; the matching direction is from left to right. Experiments show that for Chinese, reverse maximum matching method is more effective than the maximum matching method.

(2) Feature representation. Text feature representation is the textual metadata and it is divided into descriptive features (such as text name, date, size and type) and its semantic features (such as author, institutions, title and content of text). The feature representation represents documents by some feature terms, and it only needs to process these feature items during text mining so as to process unstructured text.

(3) Feature extraction. The feature extraction algorithm evaluates the feature by constructing an evaluation function, and then the feature extraction arranges the feature according to the score, which is the highest predetermined score selected. This paper takes the weight of text evidence to extract feature, the evaluation function is used to measure the difference between the probability of class and the conditional probability of given characteristic class, and its effect in the experiment is superior to the expected cross entropy.

(4) Curtailment of feature set. Feature sets are curtailed by the latent semantic indexing method and singular value decomposition (SVD) in matrix theory. The curtailment of feature set transforms the word frequency matrix to \( K \times K \) singular matrix. The basic steps are presented as follows. ① Establish the word frequent matrix. ② Do singular value analysis of word frequent matrix; decompose the frequency matrix to three matrices \( U, S \) and \( V \). The matrixes \( U \) and \( V \) are orthogonal matrices, and \( S \) is the \( K \times K \) diagonal matrix of singular values. ③ To every document \( d \), use...