Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Fatigue Crack Data Filling Based on Bayesian Theory

Liyan Fan[a] , Heyu Zhao[a] , Jincai Chang [a,b*]

[a] College of Science, North China University of Science and Technology, Hebei Tangshan 063200, China

[b] Hebei Key Laboratory of Science and Application, Hebei Tangshan 063210, China

ARTICLE INFO

Article history:
Received 15 February 2021
Accepted 20 April 2021
Available online 21 April 2021

Keywords:
Missing data
Bayesian theory
Naive bayes
padding data

ABSTRACT

China has ushered in the era of big data in an all-round way, at present. Under the background of the new era, the ability of data information collection and integration is constantly improving, making contributions to the decision-making management and development progress in various fields of society, and the quality of data also fundamentally determines the reliability of data. If the quality of data is low, it will lead to the data can not be effectively used, and even damage the availability of information systems. Therefore, the research on data quality has also received widespread attention. Among them, incomplete data and missing data is a typical problem of low data quality. This paper is based on the Bayesian theory of systematic learning to fill the missing data, using naive Bayesian theory to fill the missing data to get a complete data set, and then compare the filled data with the original data set to verify the reliability of the model.

1. Introduction

In the 1840s, Thomas Bayes, a British mathematician and reasoning statistician, first put forward Bayes theorem in his book 《On the solution of the problem of opportunity》. Because this is his posthumous work, he did not elaborate Bayes theory in more detail. Later, after sorting out and doing in-depth research on his posthumous work, other people began to study it, so that Bayesian theory is known by more people, and has been applied in many aspects. It has important applications in data processing such as data prediction, data filling and statistical inference, statistical decision-making and other statistical data.

In China, Bayesian theory is mainly used in statistical analysis, probability prediction and parameter estimation. With the in-depth study of Bayesian theory by scholars, Bayesian theory has played an important role in many aspects such as traditional Chinese medicine, railway construction and earthquake warning in recent years, rather than just limited to data statistics. At present, many people have tried to apply Bayesian theory to the processing of missing data on the basis of Bayesian theory research, and this method has also begun to become more extensive. Although the results are less at present, there has been a great breakthrough compared with before.

Bayesian theory is a complete theoretical system, and the main applications in missing data processing are naive Bayesian, Bayesian network and approximate Bayesian theory. Duan Jing introduced the naive bayes theory and its application in "Research on classification and application of naive Bayes"; Zou Wei and Wang Huijin used the combination of naive Bayes algorithm and EM algorithm to fill missing data in "EM missing data filling algorithm based on Naive Bayes"; Gong Yishan and Dong Chen used Bayesian network to repair missing data earlier; Li Zhongbo et al. Used naive bayes theory to predict protein purification, and the results improved the accuracy significantly; Oriole in the "approximate Bayesian method and its application research" introduces the approximate Bayesian method in detail, and uses it for parameter estimation, compared with the traditional method highlights the characteristics of simple and efficient.

2. Missing data

Missing data is a typical data problem in noise data. If there are errors, missing data or data which are different from other data and obviously do not conform to the data rules, these data are called noise data. According to the different data problems, noise data can be divided into a variety of situations, but this paper mainly studies missing data. Lack of data in various research areas will have a serious impact on the process and results of data mining. First of all, the system will show obvious instability because of the lack of essential data in the data set, at the same time, it will make the data

* Corresponding author.
E-mail addresses: jincai@ncst.edu.cn (J. Chang)

set without errors can not be used accurately. Secondly, the missing part of the data set will make the data processing process into confusion and lead to inaccurate results, which will directly affect the data model. In order to avoid all kinds of disadvantages caused by missing data, the processing of missing data has become an indispensable step in the era of big data. In view of the research on data missing, foreign countries have proposed approximate replacement method, neural network, stochastic regression analysis and Bayesian theory, while domestic data missing processing is mostly applied in specific cases, such as banking, insurance and so on.

2.1 Overview and processing significance of missing data

Definition of missing data: If there is no record of one or some attribute values in the data set, the data set is a data set with missing data. From the missing mechanism, the missing data is divided into completely missing at random (MCAR), missing at random (MAR), and missing at random (NMAR). If there are no missing attributes or variables in the dataset, they are complete variables, otherwise they are incomplete variables.

When the data set is missing, the results of the experiment will be greatly affected. If the data is the data of the middle process of the experiment, the experiment may even fail to proceed or the desired result cannot be obtained. In the era of big data today everyone in every profession is dealing with data almost every day, so how to deal with missing data is a realistic and urgent problem.

2.2 Treatment method of missing data

There are mainly five methods commonly used to deal with missing data:

1. Deletion method: it is divided into list deletion method and paired deletion method. It is the simplest and most direct method, but the efficiency is very low. If the data sample is large enough and the missing value is small, it can be used, otherwise more data will be lost, resulting in greater error in mining effect or statistical results.

2. Constant Completion Method: The person who uses the data is required to have a deep understanding of all aspects of the meaning of the data, and then use the constant to replace the missing value through past experience and actuality, and it is required to be familiar with the source of the data and understand the root cause of the data, otherwise it will be easy Introduce more noisy data, causing greater errors.

3. Statistical completion method: that is, through the analysis of the known complete data, the statistical information of the data set is obtained to predict the missing data. The average value completion method is commonly used, but for large data sets, but for large data sets, because too much median will produce more peak distribution, so this method will also produce serious misleading.

4. Simple value completion method: that is, the missing value is obtained by substituting simple models and calculation formulas into the data. This method is less efficient and the process of finding models and calculation formulas is more complicated.

5. Complementary method of complex estimates: that is, to estimate missing data by using the existing complete data set to establish a prediction model for missing data. This method is currently the most scientific, accurate and complex method to complete missing data. The Bayesian prediction method is one of them.

3. Bayesian theory

3.1 Bayes' theorem

Bayes' theorem was first proposed by the British scholar Thomas Bayes in his article "On the Solution to the Problem of Opportunity" in 1736. It is also called Bayesian inference. It is used to solve two random events. The formula for solving the probability problem between time. The problem of Bayesian reasoning is the problem of conditional probability reasoning. The research on conditional probability is a cornerstone of people's understanding of probability information, and it guides people to carry out effective learning and judgment. The Bayesian formula is aimed at finding the probability of two random events. If the Bayesian theory is applied to the data set to fill in the missing data, it is necessary to conduct a deeper analysis of the Bayesian formula and use the analysis applicable to practical problems. In order to combine Bayesian theory with practical problems. In the solution of the actual problem, only the prior probability is obtained through the known conditions, and then the posterior probability is calculated based on the prior probability to obtain the desired result.

As a complete decision theory, Bayesian theory includes many classic models, such as Bayesian classifier used in data preprocessing, Bayesian network involved in artificial intelligence, Bayesian statistical method learned in mathematical statistics, etc., while the most needed in data filling is naive Bayesian. The data filling model of this paper is also established on the basis of naive Bayes.

3.2 Bayesian formula

The Bayesian formula states that if there are two random events and, the probability of the event under the condition of the event is different from the probability of the event under the condition of the event, but there is a definite relationship between the two. The sub expression is the Bayesian formula.

The basic Bayesian formula is:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \tag{1}$$

Among them:

(1) $P(A)$ is the prior probability or marginal probability of A. Because it does not consider any B factors, that is, before the B event occurs, it is called "a priori", which is a simple judgment of the probability of the A event.

(2) In the same way, we can see that $P(B)$ is the prior probability or marginal probability of B, also known as the standardized constant.

(3) $P(A \mid B)$ is the conditional probability of B after the occurrence of A. After the occurrence of B, our re-evaluation of the probability of is also called the posterior probability of A.

(4) $P(B \mid A)$ is the conditional probability B of A after occurs, and it is also called the posterior probability of B.

$P(B \mid A) / P(A)$ is called the "possibility function", which is an adjustment factor that makes the estimated probability closer to the true probability.

3.3 Naive bayes theory

Naive bayes theory is simply Bayesian theory. When using

general Bayesian theory to solve problems, if the solution process is more complicated and the relationship between the attributes cannot be determined, it is assumed that the attributes are independent of each other and use calculations. The relatively simple naive bayes theory is calculated.

Next, we will quote the basic formula of naive bayes theory in data filling. $c_1, c_2, ..., c_L$ is the classification of X the data set, for any attribute of the data set, there is

$$P(c_k \mid X) = \frac{P(c_k)P(X \mid c_k)}{\sum_{r=1}^{L} P(c_r)P(X \mid c_r)}, k = 1, 2, ..., L \quad (2)$$

The prior probability is $P(c_k)$, and the posterior probability is $P(c_k \mid X)$. $P(c_k)$ can be calculated from the original data set, but if $P(X \mid c_k)$ is calculated according to the traditional Bayesian method, it will take too much time and effort in the calculation process, the naive Bayesian method assumes that the attributes are independent of each other, so it is relatively easy to calculate $P(X \mid c_k)$.

Suppose there are a total of M attributes, and classify them as $X = (X_1, X_2, ..., X_M)$, one of which X_j is divided into c_k, $k = 1, ..., L$, then the following formula holds:

$$P(X \mid c_k) = \prod_{j=1}^{M-1} P(X_j \mid c_k) \quad (3)$$

The specific process applied in data filling is as follows:

Given a data set with N records and M attributes $X_1, X_2, ..., X_M$, L_i represents the number of categories of attribute X_i, N_i represents the number of records containing known X_i attributes, and N_{ik} is when X_i is equal to it The number of records in the k-th category c_{ik}, $N_{jr|ik}$ is when X_j is equal to the number of records in its r-th category c_{jr}, and when $X_i = c_{ik}, j \neq i$.

(1) Calculate the prior probability of each attribute

$$P(X_i \mid c_{ik}) = \frac{N_{ik}}{N_i}, i = 1, 2, ..., L_i \quad (4)$$

(2) When $X_i = c_{ik}$, the conditional probability of X_j

$$P(X_j = c_{jr}) = \frac{N_{jr|ik}}{N_{ik}}, j = 1, 2, ..., M; j \neq i; r = 1, 2, ..., L_j \quad (5)$$

(3) If is missing an attribute value, and this attribute value belongs to , let be the index set of all non-missing values in records, and calculate the posterior probability. Here the value does not need to be calculated, because it can be reduced during the calculation process.

$$P(X_i = c_{ik} \mid X_j) = \frac{P(X_i = c_{ik})}{P(X_j)} \prod_{j \in J} P(X_j = c_{jr} \mid X_i = c_{ik}) \quad (6)$$

where $k = 1, 2, ..., L_i$.

4. Data analysis

4.1 Data preprocessing

After getting the fatigue crack damage experiment data of aviation aluminum alloy structure, we screened part of the complete data set. In order to facilitate the experiment, some initial data with low reference value and the last row of fracture data were removed, because the revolution is recorded in real time, there is no strict time limit. The data selected in this paper increase with the rule of 200 and 500 each time, and considering that the missing data in the actual problem, if it is revolution, then the filling will be relatively simple and will not have much impact on the experiment, so the next data filling process in this paper will directly process the tensile length of the crack without considering the influence of revolution. The following are the two databases used in the experiment after processing.

Tab. 1.This is a set of experimental data of fatigue crack damage. The data before 12500 turns are all 0. After 19024 turns, the thin plate breaks. The initial crack lengths are 0.6, 0.9, 0.9 and 0.9 respectively.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
12500	0	0	0	0
13000	0.6	0.9	0.9	0.9
13200	0.6	0.9	0.9	1
13400	0.7	1	1.1	1
13600	0.8	1	1.1	1.05
13800	0.8	1.2	1.2	1.05
14000	0.8	1.3	1.2	1.05
14200	0.8	1.3	1.3	1.1
14400	0.9	1.3	1.3	1.3
14600	1.5	1.5	1.5	1.5
14800	1.5	1.5	1.6	1.5
15000	1.5	1.5	1.6	1.5
15200	1.7	1.8	1.8	1.9
15400	1.8	2	2	2
15600	1.9	2.1	2	2.1
15800	2	2.1	2.2	2.1
16000	2.1	2.2	2.2	2.2
16200	2.2	2.2	2.4	2.3
16400	2.2	2.3	2.4	2.4
16600	2.3	2.4	2.4	2.4
16800	2.3	2.5	2.5	2.5
17000	2.4	2.5	2.5	2.5
17100	2.5	2.6	2.6	2.6
17200	2.5	2.7	2.7	2.6
17300	2.6	2.8	2.9	2.7
17400	2.6	3	3	2.8
17500	2.7	3	3.1	2.9
17600	2.8	3.1	3.1	2.9
17700	2.8	3.2	3.2	3
17800	2.9	3.2	3.2	3
17900	3	3.3	3.3	3.1
18000	3	3.4	3.3	3.1
18100	3.1	3.6	3.4	3.2
18200	3.2	3.6	3.5	3.3
18300	3.4	3.8	3.7	3.4
18400	3.5	4	4	3.5
18500	3.7	4.2	4.4	3.7
18600	3.7	4.5	4.7	3.8
18700	3.8	4.7	5	4
18800	4	4.7	5	4.2
18825	4.3	4.7	5	4.6
18850	4.5	4.7	5	5
18875	4.7	4.7	5	5.2
18900	5	4.7	5	5.6
18925	5.2	4.7	5	5.9
18950	5.5	4.7	5	6.2
18975	5.7	4.7	5	6.6
19000	6	4.7	5	7
19024	crack			

Tab. 2. This is another set of experimental data of fatigue crack damage. All data before 14500 revolutions are 0,and there is no data record after 34281 revolutions. The initial crack lengths were 0.1, 0.1, 0 and 0.1, respectively..

Number of revolutions	A/mm	B/mm	C/mm	D/mm
14500	0	0	0	0
15000	0.1	0.1	0	0.1
15500	0.1	0.1	0	0.1
16000	0.2	0.2	0.1	0.1
16500	0.2	0.2	0.1	0.1
17000	0.3	0.3	0.2	0.2
17500	0.4	0.4	0.2	0.2
18000	0.4	0.4	0.2	0.2
18500	0.5	0.4	0.3	0.2
19000	0.5	0.5	0.4	0.3
19500	0.6	0.5	0.4	0.3
20000	0.6	0.5	0.4	0.3
20500	0.6	0.6	0.5	0.3
21000	0.7	0.7	0.5	0.4
21500	0.7	0.8	0.6	0.4
22000	0.8	0.9	0.6	0.4
22250	0.9	0.9	0.6	0.5
22500	1	1	0.7	0.5
22750	1.1	1	0.7	0.6
23000	1.1	1.1	0.8	0.6
23250	1.2	1.2	0.8	0.7
23500	1.3	1.2	0.9	0.7
23750	1.4	1.3	0.9	0.8
24000	1.4	1.3	1	0.8
24250	1.5	1.3	1	0.8
24500	1.5	1.4	1	0.9
24750	1.6	1.4	1.1	0.9
25000	1.6	1.5	1.1	1
25250	1.7	1.5	1.2	1
25500	1.7	1.6	1.2	1.1
25750	1.8	1.6	1.2	1.1
26000	1.8	1.6	1.2	1.2
26250	1.9	1.7	1.3	1.3
26500	1.9	1.7	1.4	1.3
26750	1.9	1.8	1.5	1.4
27000	2	1.9	1.6	1.4
27250	2	2	1.7	1.5
27500	2	2.1	1.8	1.6
27750	2.1	2.2	1.9	1.7
28000	2.2	2.3	1.9	1.8
28250	2.3	2.3	2	1.8
28500	2.4	2.4	2.1	1.8
28750	2.5	2.5	2.1	1.8
29000	2.6	2.6	2.2	1.9
29250	2.7	2.7	2.2	1.9
29500	2.8	2.8	2.3	1.9
29750	2.9	2.9	2.4	2
30000	3	3	2.5	2
30250	3	3.1	2.6	2.1
30500	3.1	3.2	2.7	2.2
30750	3.2	3.3	2.8	2.3
31000	3.3	3.4	2.9	2.4
31250	3.4	3.5	3	2.5
31500	3.5	3.6	3.1	2.6
31750	3.7	3.8	3.2	2.8
32000	3.9	4	3.5	3
32250	4.1	4.4	3.8	3.2
32500	4.4	5	4.1	3.4
32750	4.6	5.5	4.5	3.7
33000	4.8	6	5	4
33250	5.1	6	5	4.2
33500	5.5	6	5	4.5
33750	6	6	5	5
34000	6.8	6	5	5.8
34250	9	6	5	8.5
34281	17.7	6	5	17.3

Tab. 3. For the plate specimen with two holes, the two sides of the crack of the first data are A, B, C and D respectively, the initial crack lengths are 0.6, 0.9, 0.9 and 0.9respectively, and the number of turns increases with 200 in turn. The crack width of the two holes increases irregularly with the increase of turns until fracture, and the data is only collected before fracture.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
13000	0.6	0.9	0.9	0.9
13200	0.6	0.9	0.9	1
13400	0.7	1	1.1	1
13600	0.8	1	1.1	1.05
13800	0.8	1.2	1.2	1.05
14000	0.8	1.3	1.2	1.05
14200	0.8	1.3	1.3	1.1
14400	0.9	1.3	1.3	1.3
14600	1.5	1.5	1.5	1.5
14800	1.5	1.5	1.6	1.5
15000	1.5	1.5	1.6	1.5
15200	1.7	1.8	1.8	1.9
15400	1.8	2	2	2
15600	1.9	2.1	2	2.1
15800	2	2.1	2.2	2.1
16000	2.1	2.2	2.2	2.2
16200	2.2	2.2	2.4	2.3
16400	2.2	2.3	2.4	2.4
16600	2.3	2.4	2.4	2.4
16800	2.3	2.5	2.5	2.5
17000	2.4	2.5	2.5	2.5
17200	2.5	2.7	2.7	2.6
17400	2.6	3	3	2.8
17600	2.8	3.1	3.1	2.9
17800	2.9	3.2	3.2	3
18000	3	3.4	3.3	3.1
18200	3.2	3.6	3.5	3.3
18400	3.5	4	4	3.5
18600	3.7	4.5	4.7	3.8
18800	4	4.7	5	4.2

After sorting the database, it contains 30 records and 4 attribute values. The attribute values are A, B, C, and D respectively. Among them, there are missing data in the four attribute values.

Tab. 4. For the plate specimen with two holes, the two sides of the crack of the first data are A, B, C and D respectively, the initial crack lengths are 0.1, 0.1, 0 and 0.1respectively, and the number of turns increases with 500 in turn. The crack width of the two holes increases irregularly with the increase of turns until fracture, and the data is only collected before fracture.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
15000	0.1	0.1	0	0.1
15500	0.1	0.1	0	0.1
16000	0.2	0.2	0.1	0.1
16500	0.2	0.2	0.1	0.1
17000	0.3	0.3	0.2	0.2
17500	0.4	0.4	0.2	0.2
18000	0.4	0.4	0.2	0.2
18500	0.5	0.4	0.3	0.2
19000	0.5	0.5	0.4	0.3
19500	0.6	0.5	0.4	0.3
20000	0.6	0.5	0.4	0.3
20500	0.6	0.6	0.5	0.3
21000	0.7	0.7	0.5	0.4
21500	0.7	0.8	0.6	0.4
22000	0.8	0.9	0.6	0.4
22250	0.9	0.9	0.6	0.5
22500	1	1	0.7	0.5
23000	1.1	1.1	0.8	0.6
23500	1.3	1.2	0.9	0.7
24000	1.4	1.3	1	0.8
24500	1.5	1.4	1	0.9
25000	1.6	1.5	1.1	1
25500	1.7	1.6	1.2	1.1
26000	1.8	1.6	1.2	1.2

26500	1.9	1.7	1.4	1.3
27000	2	1.9	1.6	1.4
27500	2	2.1	1.8	1.6
28000	2.2	2.3	1.9	1.8
28500	2.4	2.4	2.1	1.8
29000	2.6	2.6	2.2	1.9
29500	2.8	2.8	2.3	1.9
30000	3	3	2.5	2
30500	3.1	3.2	2.7	2.2
31000	3.3	3.4	2.9	2.4
31500	3.5	3.6	3.1	2.6
32000	3.9	4	3.5	3
32500	4.4	5	4.1	3.4

Because the number of turns has no effect on the experiment, the interval of the number of turns between the second group of data and the first group of data is different. The number of turns increases by 500 in turn. The cracks of the two holes increase irregularly with the increase of the number of turns until fracture. The data is only collected before fracture. There are 36 groups of data in this data set.

For the first data, the values in the four attribute values are divided into different intervals. The data in $[0,1]$ in A is recorded as A_1, the data in $[1.1,2]$ is recorded as A_2, and the data in $[2.1,3]$ is recorded as A_5, the data in is denoted as , the data in is denoted as , and similarly, we can see that there are five cases of B_1, B_2, B_3, B_4, B_5 in attribute B, and in attribute C there are five cases of C_1, C_2, C_3, C_4, C_5. There are five cases of D_1, D_2, D_3, D_4, D_5 in the attribute D, so the table is changed to the following form:

Tab. 5. For the data belonging to [0,1] in a, all are replaced with A1, and the data belonging to [1.1,2] are replaced with A2, and so on, all the data in a, B, C and D are replaced with tags related to attribute values a, B, C and D, so as to calculate the probability of each interval.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
13000	A1	B1	C1	D1
13200	A1	B1	C1	D1
13400	A1	B1	C2	D1
13600	A1	B1	C2	D2
13800	A1	B2	C2	D2
14000	A1	B2	C2	D2
14200	A1	B2	C2	D2
14400	A1	B2	C2	D2
14600	A2	B2	C2	D2
14800	A2	B2	C2	D2
15000	A2	B2	C2	D2
15200	A2	B2	C2	D2
15400	A2	B2	C2	D2
15600	A2	B3	C2	D3
15800	A2	B3	C3	D3
16000	A3	B3	C3	D3
16200	A3	B3	C3	D3
16400	A3	B3	C3	D3
16600	A3	B3	C3	D3
16800	A3	B3	C3	D3
17000	A3	B3	C3	D3
17200	A3	B3	C3	D3
17400	A3	B3	C3	D3
17600	A3	B4	C4	D3
17800	A3	B4	C4	D3
18000	A3	B4	C4	D4
18200	A4	B4	C4	D4
18400	A4	B4	C4	D4
18600	A4	B5	C5	D4
18800	A4	B5	C5	D5

Same as the previous data classification principle, the four attribute values are divided according to the interval, and each value is represented by the corresponding symbol to replace the original specific number, which is used to calculate the proportion of each interval in each attribute for future calculation.

Tab. 6. For the data belonging to [0,1] in a, all are replaced with A1, and the data belonging to [1.1,2] are replaced with A2, and so on, all the data in a, B, C and D are replaced with tags related to attribute values a, B, C and D, so as to calculate the probability of each interval.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
15000	A1	B1	C1	D1
15500	A1	B1	C1	D1
16000	A1	B1	C1	D1
16500	A1	B1	C1	D1
17000	A1	B1	C1	D1
17500	A1	B1	C1	D1
18000	A1	B1	C1	D1
18500	A1	B1	C1	D1
19000	A1	B1	C1	D1
19500	A1	B1	C1	D1
20000	A1	B1	C1	D1
20500	A1	B1	C1	D1
21000	A1	B1	C1	D1
21500	A1	B1	C1	D1
22000	A1	B1	C1	D1
22500	A1	B1	C1	D1
23000	A2	B2	C1	D1
23500	A2	B2	C1	D1
24000	A2	B2	C1	D1
24500	A2	B2	C1	D1
25000	A2	B2	C2	D1
25500	A2	B2	C2	D2
26000	A2	B2	C2	D2
26500	A2	B2	C2	D2
27000	A2	B2	C2	D2
27500	A2	B3	C2	D2
28000	A3	B3	C2	D2
28500	A3	B3	C3	D2
29000	A3	B3	C3	D2
29500	A3	B3	C3	D2
30000	A3	B3	C3	D2
30500	A4	B4	C3	D3
31000	A4	B4	C3	D3
31500	A4	B5	C4	D3
32000	A4	B5	C4	D3
32500	A5	B5	C5	D3

Combining the fatigue crack damage data of aviation aluminum alloy structure, we plan to find a reasonable and practical algorithm for data filling, the missing data used in the experiment are all the data sets obtained by manually removing the initial data, and the missing data sets used in this article are as follows :

Tab. 7. A random deletion operation is performed on a few rows of the first set of raw data. The random deletion here means that one of the four values in a row of data is randomly deleted in the original data and then marked. The following table shows four lines of data randomly deleted.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
14000	A1	?	C2	D2
15600	A2	B3	?	D3
17200	?	B3	C3	D3
18600	A4	B5	C5	?

The following table shows the data sets with missing data obtained from the second group of data :

Tab. 8.Randomly delete several lines of the second group of original data. The random deletion here is the same as the random deletion in the first group. The

following table shows the four lines of data after random deletion.

Number of revolutions	A/mm	B/mm	C/mm	D/mm
17000	A1	?	C1	D1
23000	A2	B2	?	D1
28000	?	B3	C2	D2
32500	A5	B5	C5	?

5. Data filling

5.1 Introducing Bayes' Theorem

First, substitute the Bayesian formula required for data filling. Then divide the sample data into $A_1,...,A_L$ and other parts. For any event in the sample space, there are

$$P(A_k \mid \theta) = \frac{P(A_k)P(\theta \mid A_k)}{\sum_{i=1}^{L} P(A_i)P(\theta \mid A_i)}, k = 1,2,...,L \tag{7}$$

$P(A_k)$ is the prior probability, and $P(A_k \mid \theta)$ is the posterior probability. $P(A_k)$ is easy to calculate from the data, and assuming that the attributes are independent of each other, so $P(A_k \mid \theta)$ can also be calculated. Assuming that there are N attributes in the sample, and θ is divided into $\theta = (\theta_1, \theta_2, ..., \theta_n)$, any θ_n is divided into $A_k, k = 1,2,...,L$, the following formula can be obtained

$$P(\theta \mid A_k) = \prod_{j=1}^{M-1} P(\theta_j \mid A_k) \tag{8}$$

5.2 Calculating the probability of missing values

This part is implemented by Python. First, input the data used, and then preprocess the data to calculate the probability of each attribute value and the conditional probability between it and other attribute values. Then, input other known values of the missing value line to get the probability of each interval of the missing value. The flow chart is shown in Figure 1：

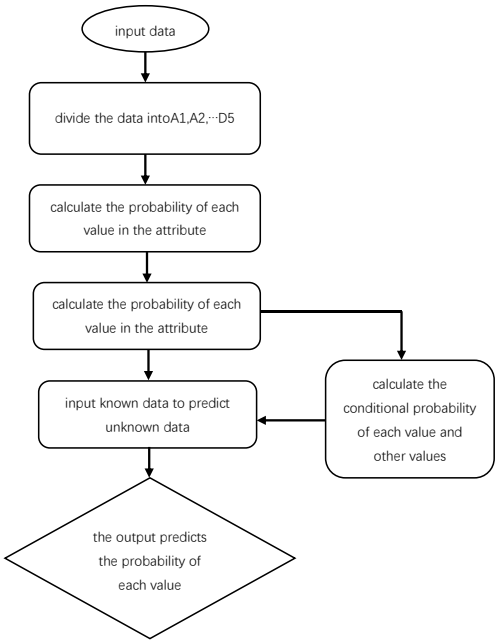

Fig. 1. After importing data, the data is preprocessed, which divides the data into intervals and calculates the probability of each value and the conditional probability of each value and other values. Finally, the unknown data is predicted by inputting known data and the probability of predicting each value is obtained.

First, calculate the probability of each interval that needs to be used to fill in the missing data among the four attribute values, namely:

Tab. 9. The data is divided into intervals, and the original data is replaced by the name of the interval after division, and the proportion of each interval of the first data in the total interval length is calculated.

$P(A_1)$	4/15	$P(B_1)$	2/15	$P(C_1)$	1/15	$P(D_1)$	1/10
$P(A_2)$	7/30	$P(B_2)$	3/10	$P(C_2)$	2/5	$P(D_2)$	1/3
$P(A_3)$	11/30	$P(B_3)$	1/3	$P(C_3)$	3/10	$P(D_3)$	2/5
$P(A_4)$	2/15	$P(B_4)$	1/6	$P(C_4)$	1/6	$P(D_4)$	2/15
$P(A_5)$	0	$P(B_5)$	1/15	$P(C_5)$	1/15	$P(D_5)$	1/30

Tab. 10. The data is divided into intervals, and the original data is replaced by the name of the interval after division, and the proportion of each interval of the second data in the total interval length is calculated.

$P(A_1)$	4/9	$P(B_1)$	4/9	$P(C_1)$	5/9	$P(D_1)$	7/12
$P(A_2)$	5/18	$P(B_2)$	1/4	$P(C_2)$	7/36	$P(D_2)$	5/18
$P(A_3)$	5/36	$P(B_3)$	1/6	$P(C_3)$	1/6	$P(D_3)$	5/36
$P(A_4)$	1/9	$P(B_4)$	1/18	$P(C_4)$	1/18	$P(D_4)$	0
$P(A_5)$	1/36	$P(B_5)$	1/12	$P(C_5)$	1/36	$P(D_5)$	0

Then combine the prior probability to calculate the conditional probability between each attribute;

Finally, calculate the posterior probabilities of the missing array of the first set of data. Since the attributes of A, B, C,and D are considered to be independent of each other, the naive Bayes method can be used directly, and then:

Fill in the missing data in the first row:

$$\begin{aligned} &P(B_1 \mid A_1, C_2, D_2) \\ &= \frac{P(B_1)P(A_1, C_2, D_2 \mid B_1)}{P(A_1, C_2, D_2)} \\ &= \frac{P(B_1)P(A_1 \mid B_1)P(C_2 \mid B_1)P(D_2 \mid B_1)}{P(A_1, C_2, D_2)} \end{aligned} \tag{9}$$

$$\begin{aligned} &P(B_2 \mid A_1, C_2, D_2) \\ &= \frac{P(B_2)P(A_1, C_2, D_2 \mid B_2)}{P(A_1, C_2, D_2)} \\ &= \frac{P(B_2)P(A_1 \mid B_2)P(C_2 \mid B_2)P(D_2 \mid B_2)}{P(A_1, C_2, D_2)} \end{aligned} \tag{10}$$

$$\begin{aligned} &P(B_3 \mid A_1, C_2, D_2) \\ &= \frac{P(B_3)P(A_1, C_2, D_2 \mid B_3)}{P(A_1, C_2, D_2)} \\ &= \frac{P(B_3)P(A_1 \mid B_3)P(C_2 \mid B_3)P(D_2 \mid B_3)}{P(A_1, C_2, D_2)} \end{aligned} \tag{11}$$

Fill in the missing data in the second row:

$$P(C_1 \mid A_2, B_3, D_3) = \frac{P(C_1)P(A_2, B_3, D_3 \mid C_1)}{P(A_2, B_3, D_3)} = \frac{P(C_1)P(A_2 \mid C_1)P(B_3 \mid C_1)P(D_3 \mid C_1)}{P(A_2, B_3, D_3)} \quad (12)$$

$$P(C_2 \mid A_2, B_3, D_3) = \frac{P(C_2)P(A_2, B_3, D_3 \mid C_2)}{P(A_2, B_3, D_3)} = \frac{P(C_2)P(A_2 \mid C_2)P(B_3 \mid C_2)P(D_3 \mid C_2)}{P(A_2, B_3, D_3)} \quad (13)$$

$$P(C_3 \mid A_2, B_3, D_3) = \frac{P(C_3)P(A_2, B_3, D_3 \mid C_3)}{P(A_2, B_3, D_3)} = \frac{P(C_3)P(A_2 \mid C_3)P(B_3 \mid C_3)P(D_3 \mid C_3)}{P(A_2, B_3, D_3)} \quad (14)$$

$$P(C_4 \mid A_2, B_3, D_3) = \frac{P(C_4)P(A_2, B_3, D_3 \mid C_4)}{P(A_2, B_3, D_3)} = \frac{P(C_4)P(A_2 \mid C_4)P(B_3 \mid C_4)P(D_3 \mid C_4)}{P(A_2, B_3, D_3)} \quad (15)$$

Fill in the missing data in the third row:

$$P(A_1 \mid B_3, C_3, D_3) = \frac{P(A_1)P(B_3, C_3, D_3 \mid A_1)}{P(B_3, C_3, D_3)} = \frac{P(A_1)P(B_3 \mid A_1)P(C_3 \mid A_1)P(D_3 \mid A_1)}{P(B_3, C_3, D_3)} \quad (16)$$

$$P(A_2 \mid B_3, C_3, D_3) = \frac{P(A_2)P(B_3, C_3, D_3 \mid A_2)}{P(B_3, C_3, D_3)} = \frac{P(A_2)P(B_3 \mid A_2)P(C_3 \mid A_2)P(D_3 \mid A_2)}{P(B_3, C_3, D_3)} \quad (17)$$

$$P(A_3 \mid B_3, C_3, D_3) = \frac{P(A_3)P(B_3, C_3, D_3 \mid A_3)}{P(B_3, C_3, D_3)} = \frac{P(A_3)P(B_3 \mid A_3)P(C_3 \mid A_3)P(D_3 \mid A_3)}{P(B_3, C_3, D_3)} \quad (18)$$

$$P(A_4 \mid B_3, C_3, D_3) = \frac{P(A_4)P(B_3, C_3, D_3 \mid A_4)}{P(B_3, C_3, D_3)} = \frac{P(A_4)P(B_3 \mid A_4)P(C_3 \mid A_4)P(D_3 \mid A_4)}{P(B_3, C_3, D_3)} \quad (19)$$

Fill in the missing data in the fourth row:

$$P(D_3 \mid A_4, B_5, C_5) = \frac{P(D_3)P(A_4, B_5, C_5 \mid D_3)}{P(A_4, B_5, C_5)} = \frac{P(D_3)P(A_4 \mid D_3)P(B_5 \mid D_3)P(C_5 \mid D_3)}{P(A_4, B_5, C_5)} \quad (20)$$

$$P(D_4 \mid A_4, B_5, C_5) = \frac{P(D_4)P(A_4, B_5, C_5 \mid D_4)}{P(A_4, B_5, C_5)} = \frac{P(D_4)P(A_4 \mid D_4)P(B_5 \mid D_4)P(C_5 \mid D_4)}{P(A_4, B_5, C_5)} \quad (21)$$

$$P(D_5 \mid A_4, B_5, C_5) = \frac{P(D_5)P(A_4, B_5, C_5 \mid D_5)}{P(A_4, B_5, C_5)} = \frac{P(D_5)P(A_4 \mid D_5)P(B_5 \mid D_5)P(C_5 \mid D_5)}{P(A_4, B_5, C_5)} \quad (22)$$

In the specific calculation process, the denominator value is the same in the calculation process of each missing value, that is, each numerator is different, and the denominator is the same. Then the proportion of the numerator is the proportion of the value in the whole attribute. Take the second behavior of the missing array of the first group of data as an example, because C has 5 attributes C_1, C_2, C_3, C_4, C_5, that is, the sum of all attribute probabilities of C is 1, so the value of $P(A_2, B_3, D_3)$ does not need to be calculated specifically, and it has no effect on the result. Other formulas have the same characteristics too. By implementing the calculation process in Python, the interval probability corresponding to each missing value can be obtained, after calculation, the posterior probability of the first group of missing arrays is as follows:

Tab. 11. Fill in the previous table with data vacancy, calculate the probability of all possible values of each vacancy, and compare with the original data.

Number of revolutions		14000	15600	17200	18600
A/mm	A1	—	—	0	—
	A2	—	—	0.0679	—
	A3	—	—	0.9321	—
	A4	—	—	0	—
B/mm	B1	0.1066	—	—	—
	B2	0.8934	—	—	—
	B3	0	—	—	—
	B4	0	—	—	—
	B5	0	—	—	—
C/mm	C1	—	0	—	—
	C2	—	0.0403	—	—
	C3	—	0.9597	—	—
	C4	—	0	—	—
	C5	—	0	—	—
D/mm	D1	—	—	—	0

D2	—	—	—	0
D3	—	—	—	0
D4	—	—	—	0.5531
D5	—	—	—	0.4469

As the first group of data missing arrays, the probability of each region used to fill in missing data in four attribute values is calculated first, and then the conditional probability between each attribute is calculated by combining the prior probability. Fill the missing array of the second set of data.

Fill in the missing data in the first row:

$$\begin{aligned} &P(B_1 \mid A_1, C_1, D_1) \\ &= \frac{P(B_1)P(A_1, C_1, D_1 \mid B_1)}{P(A_1, C_1, D_1)} \\ &= \frac{P(B_1)P(A_1 \mid B_1)P(C_1 \mid B_1)P(D_1 \mid B_1)}{P(A_1, C_1, D_1)} \end{aligned} \tag{23}$$

$$\begin{aligned} &P(B_2 \mid A_1, C_1, D_1) \\ &= \frac{P(B_2)P(A_1, C_1, D_1 \mid B_2)}{P(A_1, C_1, D_1)} \\ &= \frac{P(B_2)P(A_1 \mid B_2)P(C_1 \mid B_2)P(D_1 \mid B_2)}{P(A_1, C_1, D_1)} \end{aligned} \tag{24}$$

Fill in the missing data in the second row:

$$\begin{aligned} &P(C_1 \mid A_2, B_2, D_1) \\ &= \frac{P(C_1)P(A_2, B_2, D_1 \mid C_1)}{P(A_2, B_2, D_1)} \\ &= \frac{P(C_1)P(A_2 \mid C_1)P(B_2 \mid C_1)P(D_1 \mid C_1)}{P(A_2, B_2, D_1)} \end{aligned} \tag{25}$$

$$\begin{aligned} &P(C_2 \mid A_2, B_2, D_1) \\ &= \frac{P(C_2)P(A_2, B_2, D_1 \mid C_2)}{P(A_2, B_2, D_1)} \\ &= \frac{P(C_2)P(A_2 \mid C_2)P(B_2 \mid C_2)P(D_1 \mid C_2)}{P(A_2, B_2, D_1)} \end{aligned} \tag{26}$$

$$\begin{aligned} &P(C_3 \mid A_2, B_2, D_1) \\ &= \frac{P(C_3)P(A_2, B_2, D_1 \mid C_3)}{P(A_2, B_2, D_1)} \\ &= \frac{P(C_3)P(A_2 \mid C_3)P(B_2 \mid C_3)P(D_1 \mid C_3)}{P(A_2, B_2, D_1)} \end{aligned} \tag{27}$$

Fill in the missing data in the third row:

$$\begin{aligned} &P(A_1 \mid B_3, C_2, D_2) \\ &= \frac{P(A_1)P(B_3, C_2, D_2 \mid A_1)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_1)P(B_3 \mid A_1)P(C_2 \mid A_1)P(D_2 \mid A_1)}{P(B_3, C_2, D_2)} \end{aligned} \tag{28}$$

$$\begin{aligned} &P(A_2 \mid B_3, C_2, D_2) \\ &= \frac{P(A_2)P(B_3, C_2, D_2 \mid A_2)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_2)P(B_3 \mid A_2)P(C_2 \mid A_2)P(D_2 \mid A_2)}{P(B_3, C_2, D_2)} \end{aligned} \tag{29}$$

$$\begin{aligned} &P(A_3 \mid B_3, C_2, D_2) \\ &= \frac{P(A_3)P(B_3, C_2, D_2 \mid A_3)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_3)P(B_3 \mid A_3)P(C_2 \mid A_3)P(D_2 \mid A_3)}{P(B_3, C_2, D_2)} \end{aligned} \tag{30}$$

$$\begin{aligned} &P(A_3 \mid B_3, C_2, D_2) \\ &= \frac{P(A_3)P(B_3, C_2, D_2 \mid A_3)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_3)P(B_3 \mid A_3)P(C_2 \mid A_3)P(D_2 \mid A_3)}{P(B_3, C_2, D_2)} \end{aligned} \tag{31}$$

$$\begin{aligned} &P(A_4 \mid B_3, C_2, D_2) \\ &= \frac{P(A_4)P(B_3, C_2, D_2 \mid A_4)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_4)P(B_3 \mid A_4)P(C_2 \mid A_4)P(D_2 \mid A_4)}{P(B_3, C_2, D_2)} \end{aligned} \tag{32}$$

$$\begin{aligned} &P(A_5 \mid B_3, C_2, D_2) \\ &= \frac{P(A_5)P(B_3, C_2, D_2 \mid A_5)}{P(B_3, C_2, D_2)} \\ &= \frac{P(A_5)P(B_3 \mid A_5)P(C_2 \mid A_5)P(D_2 \mid A_5)}{P(B_3, C_2, D_2)} \end{aligned} \tag{33}$$

Fill in the missing data in the fourth row:

$$\begin{aligned} &P(D_2 \mid A_5, B_5, C_5) \\ &= \frac{P(D_2)P(A_5, B_5, C_5 \mid D_2)}{P(A_5, B_5, C_5)} \\ &= \frac{P(D_2)P(A_5 \mid D_2)P(B_5 \mid D_2)P(C_5 \mid D_2)}{P(A_5, B_5, C_5)} \end{aligned} \tag{34}$$

$$\begin{aligned} &P(D_3 \mid A_5, B_5, C_5) \\ &= \frac{P(D_3)P(A_5, B_5, C_5 \mid D_3)}{P(A_5, B_5, C_5)} \\ &= \frac{P(D_3)P(A_5 \mid D_3)P(B_5 \mid D_3)P(C_5 \mid D_3)}{P(A_5, B_5, C_5)} \end{aligned} \tag{35}$$

$$\begin{aligned} &P(D_4 \mid A_5, B_5, C_5) \\ &= \frac{P(D_4)P(A_5, B_5, C_5 \mid D_4)}{P(A_5, B_5, C_5)} \\ &= \frac{P(D_4)P(A_5 \mid D_4)P(B_5 \mid D_4)P(C_5 \mid D_4)}{P(A_5, B_5, C_5)} \end{aligned} \tag{36}$$

$$
\begin{aligned}
&P(D_5 \mid A_5, B_5, C_5) \\
&= \frac{P(D_5)P(A_5, B_5, C_5 \mid D_5)}{P(A_5, B_5, C_5)} \\
&= \frac{P(D_5)P(A_5 \mid D_5)P(B_5 \mid D_5)P(C_5 \mid D_5)}{P(A_5, B_5, C_5)}
\end{aligned} \tag{37}
$$

After calculation, the posterior probability of the second missing array is as follows:

Tab. 12. Fill in the previous table with data vacancy of the missing array of the second set of data , calculate the probability of all possible values of each vacancy, and compare with the original data.

Number of revolutions		17000	23000	28000	32500
A/mm	A1	—	—	0	—
	A2	—	—	0.0100	—
	A3	—	—	0.0333	—
	A4	—	—	0	—
	A5	—	—	0	—
B/mm	B1	0.4667	—	—	—
	B2	0	—	—	—
	B3	0	—	—	—
	B4	0	—	—	—
	B5	0	—	—	—
C/mm	C1	—	0.0296	—	—
	C2	—	0.0204	—	—
	C3	—	0	—	—
	C4	—	0	—	—
	C5	—	0	—	—
D/mm	D1	—	—	—	0
	D2	—	—	—	0
	D3	—	—	—	0
	D4	—	—	—	0.0333
	D5	—	—	—	0

5.3 Filling data with mean method

Because the characteristics of the data increase in turn, some people will use the method of mean to fill the data. The data in this paper can also be filled with the method of mean. Take the first group of missing arrays as an example, the missing values are B2 before and after. If we use the method of mean to fill the data, it will make the data lose diversity and the result is not accurate enough, but the probability of B1 and B2 can be obtained by Bayesian method, which can improve the diversity of data, and researchers can choose the data according to their needs. Comparing the two results, we can find that the filling results of Bayesian method are more accurate.

Compared with the original data, it can be found that the calculation method has high accuracy and can predict multiple interval probabilities. Because the amount of data used in this paper is small and the data repetition rate is low, it needs to segment the interval. If the amount of data repetition rate is high and the characteristics are obvious, it can accurately predict the specific value, which has a certain reference value for data filling.

6. Conclusion

In this paper, Bayesian theory is used to establish a missing value filling model based on the fatigue crack growth data of aviation aluminum alloy, and the accuracy of the model is verified by comparing the filled data with the original data. The model is established for the small data in this paper, and the data with large amount of data can be modified appropriately. The calculation process of the model is simple, it is easy to understand and has high accuracy, and can predict the probability of different values. The study of missing values for practical problems has a very important reference value, and can be extended to the missing sample set of three-dimensional data.

Acknowledgements

This work was supported by Ministry of education production university cooperation education project (201802305012).

References

Hruschka, J.E., Ebecken, N.F., 2002. Missing values prediction with K2. J. Intelligent Data Analysis. 6, 557-566.

Jnsson, P., Wohlin, C., 2004. An Evaluation of k-Nearest Neighbour Imputation Using Liker Data. C. 10th International Symposium on Software Metrics.IEEE. 2004, 108-118.

Zhang, T.F., Jin, X.H., Zhu, H.Q., 2021. Weapon equipment performance prediction based on improved naive Bayes. J. Journal of Military Communications College. 23, 23-27.

Batista, G.E., Monard, M.C.,2002. A study of K-nearest neighbour as an imputation method.J. His, 87,251-260.

Li, M., Zheng, Y.L., Gao, Z., Chen, D.Q., Zhan, S.L., 2021. Power material demand forecasting based on multidimensional fusion of influencing factors and Bayesian probability update. J. Logistics technology. 40, 71-76.

Ma, X., Gu, Y., 2016. Sequence sensitive multi-source perceptual data filling technology. J. Journal of software. 37, 3333-3347.

Song, Z.L., Jia, X., Guo, B., Cheng, Z.J., 2021. System residual life prediction based on Bayesian fusion and simulation. J. Systems engineering and electronic technology. 43, 1706-1713.

Wu, Z.M., 2007. Models, Methods and Theory of Scattered Data Fitting. J.

Yang, F., Hai, L., Liu, W., 2020. Event prediction of hydrogeological system based on Bayesian algorithm. J. Digital communication world. 12, 66-68.

Zou, W., Wang, H.J., 2011. EM missing data filling algorithm based on Naive Bayes. J. Microcomputer and application. 30, 75-77 + 81.

Yu, P., Yang, F.L., Kegang, L., Oyinkepreye, D.O., Yang, X.Y., 2020. Prediction and characterization technology of hydraulic units in tight reservoirs based on Bayesian inference. J. Special oil and gas reservoirs. 27, 81-87.

Qin, C., 2018. Stream data association analysis based on missing data. D. Xi'an University of Electronic Science and technology.

Gong, Y.S., Dong, C., 2010. Missing data processing based on Bayesian network. J. Journal of Shenyang University of technology. 33, 79-83.

Eckert, F.,Hyndman, R.J., Panagiotelis, A., 2020. Forecasting Swiss exports using Bayesian forecast reconciliation. J. European Journal of Operational Research.

Chen, C., 2016. Three dimensional denoising of white light interference signal based on Bayesian estimation. D. Nanchang Aeronautical University.

Wang, Y.W., Deng, L., Zheng, L.Y., Wang, Y.H., 2021. Tool residual life prediction method based on multi-channel fusion and Bayesian theory. J. Journal of mechanical engineering. 57, 214-224.

Hu, X.Z, Chen, X.X., Qian, Y.L., 2013. Research on missing data filling method in data processing. J. Journal of Hubei University of technology. 28, 82-84.

Shen, X., 2011. Research on missing data completion based on Bayesian method. D. Chongqing University.

Ren, J.M., Zhao, Y., Chen, F., LAN, S.Y., 2004. Hot deck estimation of missing data in mix design. J. China health statistics. 05, 48-51.

Shi, G.J., 2012. Research on the application of Bayesian multiple imputation method in food enterprise credit rating. D. Hunan University.

Li, X.W., Tao, Y.G., Gu, J., Liu, S.Q., Li, M., Chen, C., Li, R.Y., 2020. Distribution network net load forecasting method based on Bayesian

network. J. Electrical appliances and energy efficiency management technology. 09, 90-98.

Jin, G., 2007. A Bayes bootstrap method for synthesizing performance and life data. J. Acta Astronautica Sinica. 03, 731-734 + 771.

Bi, Z., Hou, S.L., 2020. Research on aviation material consumption prediction model based on fuzzy soft set and Bayesian. J. Ship Electronic Engineering. 40, 116-119.

Li, H., A, M.N., Li, P., Wu, M., 2010. Missing data filling algorithm based on EM and Bayesian network. J. Computer engineering and application. 46, 123-125.

Xu, J.W., 2021. Bayesian energy consumption prediction method and application of campus buildings based on monitoring data. D. Dalian University of technology.

Zhou, L.Q., Zhang, Q., 2020. Wei L., Research on big data outlier detection model based on Bayesian. J. Computer knowledge and technology. 16, 207-209.

Yao, M., Gao, G., Hu, R. Q., 2020. An improved Bayesian inversion algorithm. J. Progress in geophysics. 35, 1911-1918.

Zhou, Y., Jia, F.D., Xi, S.H., 2018. Experimental study on multi model structure recognition based on Bayesian theory. J. Journal of Hunan University (NATURAL SCIENCE EDITION). 45, 36-45.

Gong, J.H., 2017. Application of Bayesian theory in genetic probability calculation. J. Middle school biology teaching. 19, 41-43.

Kim, Y.K., Seo, J.I., 2020. Objective Bayesian Prediction of Future Record Statistics Based on the Exponentiated Gumbel Distribution: Comparison with Time-Series Prediction. J. Symmetry. 12(9).

Pang, B., Cheng, D.P., 2019. Video classification method using naive Bayes classifier. J. Journal of Wuhan Polytechnic of engineering. 31, 14-17.

Song, J., Chen, G.S., Chen, J.f., Xu, B.P., 2021. Dimension prediction of injection molded products based on feature selection and Bayesian optimization. J. Engineering plastics application. 49, 54-60.

Xu, J.W., Zhao, T.Y., Wang, P., Ma L.D., Zhang, C.Y., 2021. Bayesian energy consumption prediction model of campus buildings based on energy consumption monitoring data. J. HVAC. 51, 123-129.

Cheng, K.K., Yao, J.T, Cheng, Z.J., Dai, J.B., Song, M.M., 2021. Prediction method of pipeline corrosion depth based on correlation and Bayesian inference. Oil and gas storage and transportation. 40, 854-859.

Liyan Fan She was born in Zhang Jiakou, Hebei province, China in 1996. Now, she is a graduate student of Graduate School of North China University of science and technology, studying in math, mainly researches on numerical calculation and its application.

Heyu Zhao He was born in Shijiazhuang, Hebei province, China in 1997.Now, He is a graduate student of Graduate School of North China University of science and technology, studying in math, mainly researches on numerical calculation and its application.

Jincai Chang received his B.Sc. degree in 1996 from Ocean University of China, received his M.Sc. degree in 2005 from Yanshan University, received his Ph.D. degree in 2008 from Dalian University of technology, now he is Professor in North China University of Science and technology. His main research interests include theories and methods in mathematical modelling and scientific computation, numerical approximation and computational geometry, etc.