Journal of Science and Engineering Research
Vol. 1, No. 2, 2022, pp. 49-60
eISSN: 2957-7187

Improved Method of Filling Incomplete Information in Network Database Simulation

John W. Cheng

Tsuda University, Japan
cwljwc@tsuda.ac.jp

DOI: http://dx.doi.org/10.56828/jser.2022.1.2.4
Article history: *Received (January 29, 2022); Review Result (March 2, 2022); Accepted (May 26, 2022)*

Abstract: With the rapid development of the Internet today, database data loss in a big data environment is inevitable. How reduce data loss and supplementing incomplete data is the key to current data analysis. In this paper, a method for filling incomplete information in-network databases based on multiple regressions KNN is designed. Through coefficient adjustment, this method may effectively eliminate the effect of missing data noise on the filling result, and better solve the filling caused by missing data noise. The experimental results demonstrate that the method investigated in this paper is more efficient than the conventional method by showing that it fills in the network database's incomplete information more quickly than the traditional two filling methods. It also takes less time to detect the network database's incomplete information.

Keywords: Big data, Data filling, Data missing, Information detection

1. Introduction

At this stage, network database data loss is often an inevitable problem in the big data environment [1][2][3][4][5]. How to deal with the lack of data is the focus of current research in the field of data analysis. At this stage, a lot of data analysis relies on complete data sets, which brings some troubles. For this reason, finding an effective and feasible method to deal with these missing data is a problem that needs to be solved urgently.

At present, many scholars have carried out research on the filling of incomplete data. Simon et al., [6] proposed a method of filling missing data in different types of incomplete big data. This method mainly finds the difference between other types of indicators and certain types of indicators. In correlation, the data set is obtained, the weight coefficients are obtained, the information entropy of the initial database is calculated using related theories and experience, the lower limit of the missing data interval is determined, and the missing data is filled. However, this method takes a long time to detect missing data, which results in a low filling efficiency of incomplete information in the network database. Trzasko et al., [7] proposed a method for restoring and reconstructing lost data from distributed database users. This method mainly calculates the weight of the neighbors to obtain the filling amount of the lost data to complete the restoration and reconstruction of the lost data. However, the missing data prediction error of this method is relatively large, resulting in the low accuracy of the missing data estimate. This paper proposes a method of filling incomplete information in a network database based on multiple regressions KNN. The network database's incomplete information must first be identified and pre-processed before being estimated and filled in. This is done using the multiple regression KNN methods to determine the Euclidean distance between the target data in the network database and all the data records in the complete value data matrix. Finally, the experimental results show that the incomplete information filling method of the network database based on the multiple regressions KNN in this study is better than the traditional method and has practical application significance.

2. Filling Method of Incomplete Information in Network Database

The framework of the method of filing incomplete information of the network database based on the multiple regressions KNN studied in this paper is shown in Figure 1.

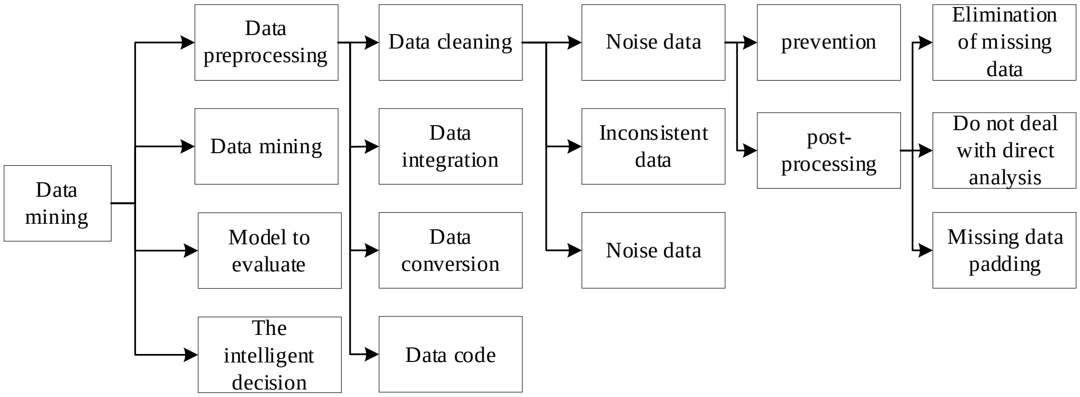

Figure 1: The incomplete information filling framework of the network database based on multiple regression KNN

Among them, data pre-processing is the basis of subsequent data mining and filling, which will directly affect the filling results. For this reason, the data is detected and pre-processed in advance. The specific content is as follows.

2.1. Incomplete information detection and preprocessing of network database

Before filling the incomplete information of the network database, the incomplete information of the network database is detected and pre-processed in advance [8][9]. Suppose there are m data objects in the sample space, and each data object contains n segmentation attributes, which is expressed as

$$\omega_k = (\sigma_k + E_k) \sum_r^{i=1} (1 - r_{ik}) \tag{1}$$

In formula (1), ω_k represents the weight of the k-th attribute value in the network database, r_{ik} represents the correlation coefficient between the i-th attribute and the k-th attribute, and E_k and σ_k respectively represent the standard deviation of the data k.

To comprehensively consider the relevance, conflict, and discrete characteristics in the network database, firstly, the gray correlation degree calculation method is used to detect all the information in the database [10][11]. Suppose that given a data area in a database, calculate the data density in the network database to obtain the detected incomplete information, expressed by the following formula

$$w = \omega_k \left(\frac{t}{n}\right) + v_i \tag{2}$$

In formula (2), t represents the current data query time, n represents the number of data samples, and v_i represents the data density of the i-th sample.

At the same time, for a network database with missing data, if a smaller interval is used to discover the relationship between the data and the data, it will increase the information entropy in the database [12]. Information entropy is a measurement parameter with the degree of system order. The larger the value increases, the more chaotic the calculation becomes. The calculation will be more consistent if the information entropy is smaller. For this reason, it is defined as

$$H(P) = -\sum_{n}^{i-1} p(w)\, log\, p(w) \tag{3}$$

In formula (3), p represents the measurement parameter.

Based on the above calculation results, using the attribute reduction algorithm of information entropy, the incomplete information in the network database is reduced [13-14]. When the collected data interval is larger, the linear correlation between the data will be reduced. The linear difference formula based on time correlation thus defined is as follows:

$$g = \sum_{2a}^{i-1} \frac{d_i}{3^i} + \eta H(P) \tag{4}$$

In formula (4), η represents the prediction error, d represents the mean value of the known data distance, and u represents the value of the linear difference sample variable. The larger the value, the more accurate the calculation.

Through the above process, the detection and pre-processing of incomplete information in the network database are completed.

2.2. Incomplete information filling based on multiple regressions KNN

Based on the above-mentioned detection and pre-processing of incomplete information in the network database, the incomplete information in the network database is filled. In this process, the multiple regression KNN methods is mainly used to fill incomplete information in the network database. The steps are as follows:

First, the network database data is initialized, and the classification interval in the network database is calculated. The expression is as follows:

$$F = \frac{e \cdot x_i + bg}{s} \tag{5}$$

The variables e, b, s, and x i in formula (5) stand for the interval value between the data and the data in the network database, the optimal classification function, the objective function of the classification, and the discriminant function of the i-th data, respectively [15].

Next, determine how far apart the target data in the network database is from each and every data record in the full value data matrix. The expression is as follows:

$$d_i = \sqrt{(z_i - o)'(z_i - o)} \tag{6}$$

In formula (6), z i stands for the i-th nearest neighbor's nearest neighbor parameter, and o for the target data.

Third, using the method outlined above, the Euclidean distance is determined. As the target data's closest neighbor, choose the data record with the shortest Euclidean distance and place it in the responsible location of the data matrix;

Fourth, from the entire value data matrix, choose the data record with the least Euclidean distance to each target data and save it in the data group [16][17];

Fifth, initialize the importance of the nearest neighbor of each target data, the expression is as follows:

$$R = \frac{F}{d_1(z_i,B)} \tag{7}$$

R stands for the relevance of neighbors, and B stands for the evaluation parameter for the significance of the data in formula (7).

Sixth, to eliminate the nearest neighbor noise of the target data, the specific judging criteria are as follows:

$$W = M/\frac{R}{d(x,x_I)} \tag{8}$$

In equation (8), M stands for the noise elimination parameter, and x I is the noise judgment result of the target data record's i-th nearest neighbor.

The target data's noise-free nearest neighbor is determined using the technique above, the noise in the closest neighbor is removed, and the missing value is determined. Based on this, the data from the original feature space is mapped to a new feature space using a suitable nonlinear function. Its expression is

$$w^* = \sum_t^{I=1} W\, \alpha_i^* y_i x_i \tag{9}$$

In formula (9), y_i represents the discriminant function, and α_i^* represents the set threshold.

Seventh, the incomplete information in the network database is estimated and filled according to the missing values based on the completion of the aforementioned data space mapping. In the process of processing, it should be noted that in most cases the component data in the database is different, that is, each row of the data in the database is different data, which is expressed as

$$\boldsymbol{Q} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{bmatrix} \tag{10}$$

The aforementioned matrix is an observation matrix where D stands for the number of columns as the number of component data and n stands for the number of rows, or sample size.

Since the fixed sum of each observation value in the data is different, to fill the accuracy, set the adjustment factor and express it as

$$f_{ij} = \frac{V}{w^* Q c^* X_{jk}} \tag{11}$$

In formula (11), X_{jk} represents the balance component of different observations, c^* represents the adjustment factor, and V represents the missing value.

After adjustment through the above process, the consistency of the component data in the database can be ensured. Finally, fill in the incomplete information of the network database, and its expression is

$$g = \sum_{k}^{i=1} w_i / x_i f_{lj} \tag{12}$$

In formula (12), x represents the value of the nearest neighbor response position, and $\sum_{k}^{i=1} w_i$ represents the missing data judgment parameter.

Repeat the previous steps until all of the network database's incomplete information has been found and filled in, then repeat the process to finish filling out the database's incomplete information.

3. Experimental Analysis

3.1. Experiment preparation

To verify the effectiveness of the incomplete information filling method of the network database based on the multiple regression KNN, an experimental comparison and analysis were carried out, and the method of filling missing data in different types of incomplete big data was proposed by Simon et al., [6]. Trzasko et al., [7] Proposed methods for restoring and reconstructing the lost data of distributed database users are compared with the method of this research. The hardware platform of this experiment is configured as CPU-INTELCOREi7-8700K3.7GHz6-Core, etc., and the software platform is based on PyCharm using sklearn and jupyter notebook. Chart drawing.

This experimental study is divided into two experiments. In experiment 1, the data missing rate in the network database is set to 5%. The missing data detection time and the accuracy of the missing data estimated value of the three methods are mainly compared with that of the network database filling time of complete information. In experiment 2, the data missing rate of the network database is about 10%. In this experiment, the prediction error and information filling time of the missing data of the three methods are mainly compared.

3.2. Missing data detection time

In experiment 1, the method of filling incomplete information in-network databases based on multiple regression KNN was compared with the method of filling missing data in different types of incomplete big data proposed by Simon et al., [6], and the method of recovering lost data from distributed database users proposed by Trzasko et al., [7]. Figure 2 shows the comparison result of the missing data detection time of the constructive method.

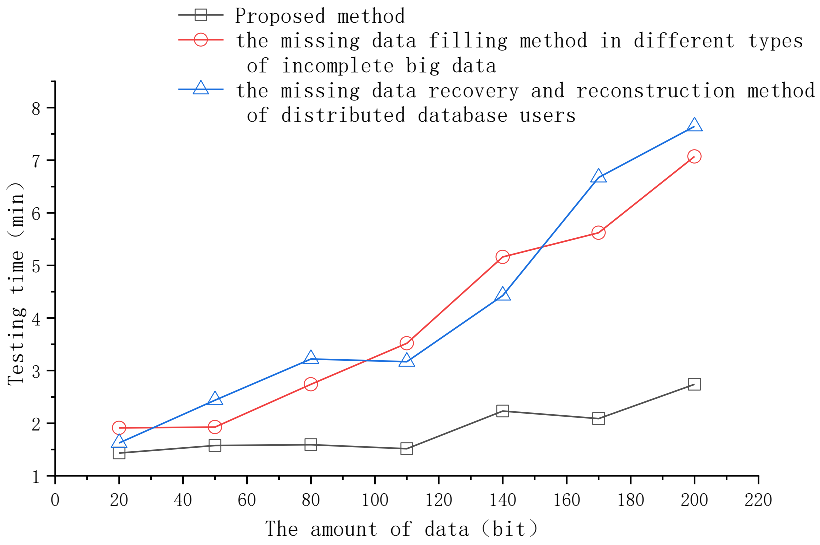

Figure 2: Comparison of missing data detection time

Analysis of the above figure shows that as the amount of data continues to increase, the missing data detection time of this research method, the missing data filling method in different types of incomplete big data, and the missing data recovery and reconstruction method of distributed database users also increases linearly. The comparison shows that the detection method of this study has less detection time than the traditional method.

3.3. Accuracy of missing data estimates

The calculation formula for the accuracy of missing data estimates is as follows:

$$ESM = \sqrt{\frac{1}{m}\sum_{m}^{i=1}(\hat{y}_{it} - y_{it})^2} \tag{13}$$

In the formula, m represents the total number of actual data, y_{it} represents the sensory data of the node at time t, and $\hat{y}_{it}$ represents the estimated value of the sensory data of the node at time t.

Taking the accuracy of the missing data estimates as the experimental indicators, a comparative analysis was carried out. The method of filling incomplete information in the network database based on the multiple regressions KNN researched in this paper was compared with the method of filling missing data in different types of incomplete big data proposed by Simon et al., [6]. The accuracy of the missing data estimated value of the distributed database user lost data recovery and reconstruction method proposed by Trzasko et al., [7] is compared and analyzed, and the comparison result is shown in Figure 3.

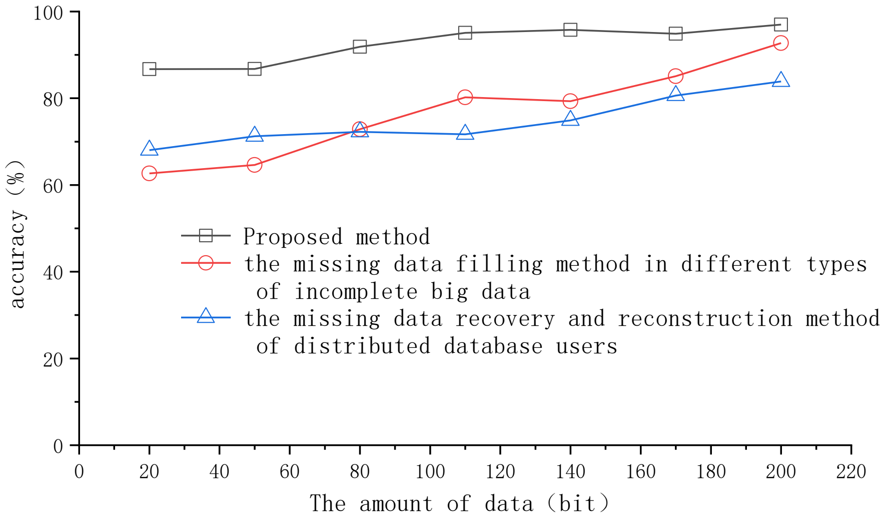

Figure 3: Comparison of accuracy of missing data estimates

Analysis of the above figure shows that when the amount of data is less than the amount of data, the missing data estimation accuracy of the incomplete information filling method of the network database based on the multiple regressions KNN in this study is higher, and the initial accuracy is slightly lower. The reason is that in the initial analysis, there is less auxiliary information. As the amount of information increases, the error decreases and gradually shows a balanced trend. After comparison, it can be seen that the filling method of this study has higher estimation accuracy than the traditional two methods.

3.4. Filling time of incomplete information when the missing rate is 5%

When the data missing rate of the network database is 5%, the incomplete information filling time of the three methods is compared, and the comparison result is shown in [Figure 4].

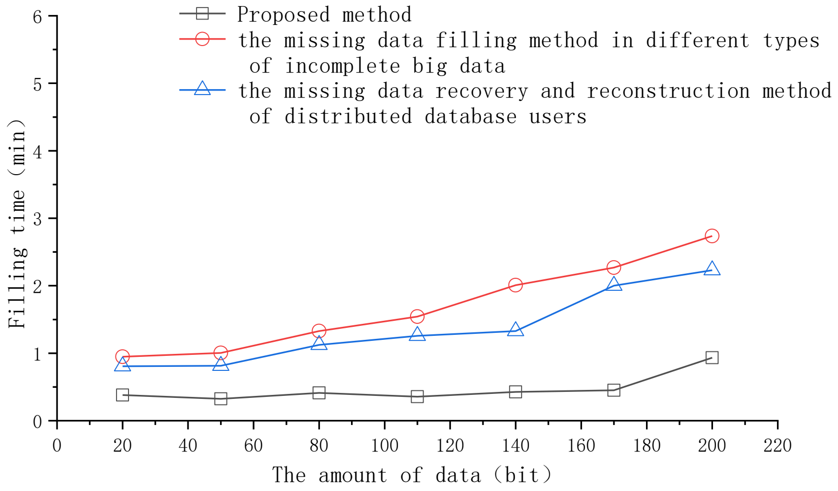

Figure 4: Filling time of incomplete information when the missing rate is 5%

The analysis of the above figure shows that the method of filling incomplete information in the network database based on multiple regressions KNN in this study takes less time to fill in incomplete information. The comparison shows that the method of this study is better than the traditional two methods. It takes less time to fill in incomplete information.

3.5. Prediction error of missing data

A comparative analysis of the missing data filling method in different categories of incomplete big data, the recovery and reconstruction method of distributed database user lost data and the prediction error of the missing data of the incomplete information filling method of the network database based on the multiple regressions KNN of this research are compared and analyzed. The comparison result is shown in [Figure 5].

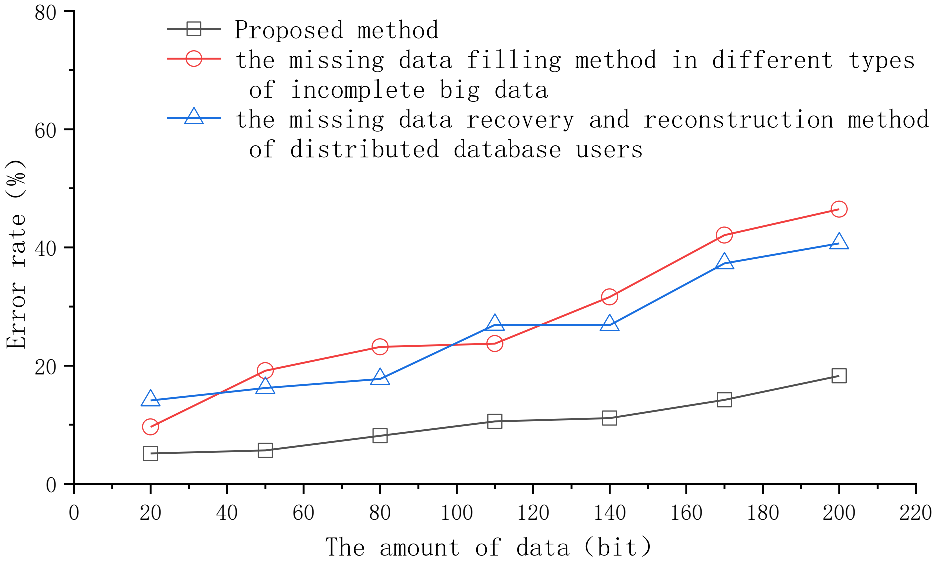

Figure 5: Forecast error of missing data

It can be seen from Figure 5 that with the increase in the amount of actual data, the prediction errors of the missing data filling method in different types of incomplete big data and the method of restoring and reconstructing the missing data of distributed database users also increase, and the magnitude of the change is relatively large. The comparison shows that there is no filling method in this study with high prediction accuracy.

3.6. Filling time of incomplete information when the missing rate is 10%

When the data missing rate of the network database is 10%, the incomplete information filling time of the three methods is compared, and the comparison result is shown in Figure 6.

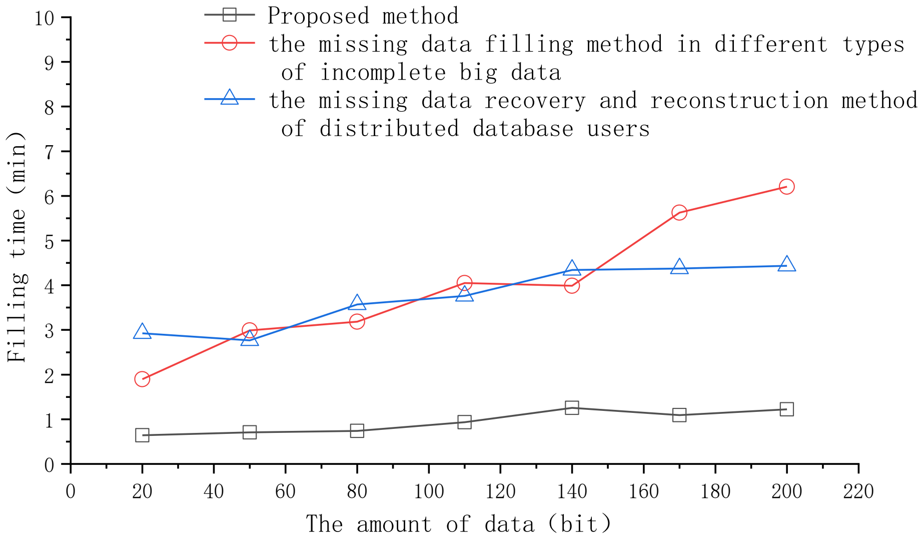

Figure 6: Filling time of incomplete information when the missing rate is 10%

From the comparison of the filling time, it can be seen that with the continuous increase of the amount of data, the filling time of the incomplete information of the three methods is also increasing. The information filling time is the shortest, and the missing data filling method in different types of incomplete big data and the distributed database user lost data recovery and reconstruction method take more time.

To sum up, the method of filling incomplete information of the network database based on multiple regressions KNN in this study has less detection time for missing data, higher accuracy of missing data estimation, low prediction error, and information filling time than the traditional two methods. Less, fully verify the effectiveness of this research method. The reason is that the filling method used in this study detects and pre-processes incomplete information in the network database in advance, and uses the multiple regression KNN methods to fill it, thereby obtaining a better filling effect and meeting the design requirements of the filling method.

4. Conclusion

Due to the noise in the missing data in the network database, the result of filling incomplete information in the network database has a large deviation. This paper proposes a method of filling incomplete information in the network database based on multiple regressions KNN. The grey correlation degree calculation method is used to detect incomplete information in the database. According to the detection result, the attribute reduction algorithm of information entropy is used to reduce incomplete information. The Euclidean distance between the target data in the network database and all other data records in the full value data matrix is calculated using the multiple regressions KNN method. The data record with the smallest Euclidean distance is chosen as the nearest neighbor of the target data, and the non-noisy nearest neighbor of the target data is assessed. Complete the elimination of nearest neighbor noise, obtain missing values, and complete the filling of incomplete information in the network database. The experimental results show that by observing the experimental results, the research method effectively reduces the missing data detection time and prediction error, shortens the time for filing incomplete information in the network database, improves the accuracy of the estimated value of the missing data, and meets the requirements of the network database. Complete information fills the requirements. However, the method of this study also has certain shortcomings. The determination of the rank of the recombination matrix obtained after the missing data is completed is not unique. How find the optimal rank still needs further research to improve the effect of filling incomplete information in the database.

References

[1] V. K. Rao & R. Caytiles. (2017). Sub-graph with set similarity in a database. *Asia-pacific Journal of Convergent Research Interchange, SoCoRI, ISSN: 2508-9080 (Print); 2671-5325 (Online)*, 3(2), 29-37. DOI:10.21742/APJCRI.2017.06.04.

[2] H. Sug. (2020). The utility of APEX in the context of database education. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, ISSN: 2508-9080 (Print); 2671-5325 (Online),* 6(2), 69-78. DOI:10.21742/apjcri.2020.02.06.

[3] M. Li, Z. Shuang, & F. Wang. (2020). Influence of internet-based social big data on personal credit reporting. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, ISSN: 2508-9080 (Print); 2671-5325 (Online),* 6(7), 39-57. DOI:10.47116/apjcri.2020.07.05.

[4] K. K. H. Kunasekaran, Y. Zheng, & W. Wang. (2020). Research on customer relationship management based on data mining. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, ISSN: 2508-9080 (Print); 2671-5325 (Online)*, 6(5), 61-77. DOI:10.21742/apjcri.2020.05.06.

[5] X. Wang. (2020). The construction of an employment and entrepreneurship service system for university students based on big data. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, ISSN: 2508-9080 (Print); 2671-5325 (Online)*, 6(10), 203-213. DOI:10.47116/apjcri.2020.10.16.

[6] G. Simon, J. A. Lee, & M. Verleysen. Double quantization forecasting method for filling missing data in the CATS time series. *IEEE International Joint Conference on Neural Networks.* IEEE.

[7] J. D. Trzasko & A. Manduca. Method for compressed sensing image reconstruction using a priori knowledge of spatial support, US.

[8] A. Marchesi, A. Bria, & C. Marrocco. (2017). The effect of mammogram preprocessing on microcalcification detection with convolutional neural networks. *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*.

[9] S. Ahmed, M. H. Yap, & M. Tan. (2020). ReCoNet: Multi-level preprocessing of chest x-rays for COVID-19 detection using convolutional neural networks.

[10] M. Si. (2019). DEA cross-efficiency ranking method based on grey correlation degree and relative entropy. *Entropy*, 21(10), 966.

[11] Y. T. Bai, X. B. Jin, & X. Y. Wang. (2020). Dynamic correlation analysis method of air pollutants in Spatio-temporal analysis. *International Journal of Environmental Research and Public Health*, 17(1).

[12] R. Miao, X. Xiang, & Q. Wu. (2020). Evaluation method of medical service system based on DEMATEL and the information entropy: A case study of hypertension diagnosis and treatment in China. *PLoS ONE*, 15(12), e0243832.

[13] H. –Y. Kim. (2011). The analysis of changes in the contract zone on pre-trial bargaining outcome under incomplete information. *Asian Journal of Law & Economics*, 2(2), 4-4.

[14] A. K. Birjandi, S. Dehmolaee, & R. Sheikh. (2020). Analysis and classification of companies on Tehran stock exchange with incomplete information. *RAIRO - Operations Research*, 55.

[15] E. N. Liberda, A. M. Zuk, & I. D. Martin. (2020). Fisher's linear discriminant function analysis and its potential utility as a tool for the assessment of health-and-wellness programs in indigenous communities. *International Journal of Environmental Research and Public Health*, 17(21), 7894.

[16] R. Zubaedah, F. Xaverius, & H. Jayawardana. (2020). Comparing Euclidean distance and nearest neighbor algorithm in an expert system for diagnosis of diabetes mellitus. *Enfermería Clínica*, 30, 374-377.

[17] S. Har-Peled & B. Raichel. (2014). Net and prune: A linear-time algorithm for Euclidean distance problems. (

This page is empty by intention.