IMPLEMENTATION OF MISSING VALUES HANDLING METHOD FOR EVALUATING THE SYSTEM/COMPONENT MAINTENANCE HISTORICAL DATA

Entin Hartini
Center for Nuclear Reactor Technology and Safety
Kawasan Puspiptek, Serpong, Tangerang Selatan, 15310
entin@batan.go.id
Diterima editor: 19 Desember 2016
Diperbaiki: 3 Februari 2017
Disetujui untuk publikasi: 16 Februari 2017

ABSTRACT

IMPLEMENTATION OF MISSING VALUES HANDLING METHOD FOR EVALUATING THE SYSTEM/COMPONENT MAINTENANCE HISTORICAL DATA. Missing values are problems in data evaluation. Missing values analysis can resolve the problem of incomplete data that is not stored properly. The missing data can reduce the precision of calculation, since the amount of information is incomplete. The purpose of this study is to implement missing values handling method for systems/components maintenance historical data evaluation in RSG GAS. Statistical methods, such as listwise deletion and mean substitution, and machine learning (KNNI), were used to determine the missing data that correspond to the systems/components maintenance historical data. Mean substitution and KNNI methods were chosen since those methods do not require the formation of predictive models for each item which is experiencing missing data. Implementation of missing data analysis on systems/components maintenance data using KNNI method results in the smallest RMSE value. The result shows that KNNI method is the best method to handle missing value compared with listwise deletion or mean substitution.

Keywords: missing value, data evaluation, alghorithm, implementation

ABSTRAK

IMPLEMENTASI METODE PENANGANAN DATA HILANG UNTUK MENGEVALUASI DATA SEJARAH PERAWATAN SISTEM/KOMPONEN. *Data hilang merupakan masalah dalam melakukan evaluasi data. Analisis data hilang dapat menyelesaikan permasalahan ketidaklengkapan data yang tidak tersimpan dengan baik. Data yang hilang akan memperkecil presisi dari perhitungan, dikarenakan jumlah informasi yang tidak lengkap. Tujuan dari penelitian ini adalah implementasi metode penanganan data hilang untuk evaluasi data sejarah perawatan sistem/komponen RSG GAS. Metodologi yang digunakan untuk menentukan data hilang yang berhubungan dengan data sejarah perawatan sistem/komponen adalah statistics, listwise deletion dan mean substitution, dan machine learning (KNNI). Metode mean substitution dan KNNI dipilih karena metode ini tidak memerlukan informasi untuk pembentukan model prediksi untuk setiap item yang mengandung data hilang. Implementasi analisis data hilang pada data perawatan sistem/komponen menggunakan metode KNNI menghasilkan nilai RMSE terkecil. Hasil ini menunjukan bahwa metode KNNI merupakan metode terbaik untuk menangani data hilang dibanding dengan listwise deletion atau mean substitution.*

Kata kunci: data hilang, evaluasi data, algoritma, implementasi

P-ISSN: 1411–240X E-ISSN: 2527-9963
Nomor: 632/AU3/P2MI-LIPI/03/2015

INTRODUCTION

Missing data is one of the problems that need to be solved for maintenance historical data evaluation. Traditional and modern methods are widely used to solve this problem. The associated variables are Missing Completely at Random, Missing at Random, Missing not at Random. Missing values can be interpreted as a data or information that is "missing" or not available for the research subject in a particular variable factor due to a non-sampling error. This happens for various reasons, such as the procedural errors of data entry, poor data storage, and so forth. Missing values in the processes can decrease the level of data accuracy. Three types of problems are usually associated with missing values ie: Loss of efficiency, complications in handling and analyzing the data and bias resulting from differences between missing and complete data [1, 2].

Currently, the maintenance historical data of systems/components RSG GAS is contained incomplete data (missing values). Therefore, it is necessary to handle missing data in order to obtain a complete data. There are several ways to handle missing values. The easiest way is to delete the data row that has missing values on attributes before it is processed (listwise deletion). Another method is imputation. It fills missing values with a value derived from the processing rows of data that are provided by a specific algorithm (mean substitution) and a method that requires a distribution function of the estimated data (maximum expectation)[2–7]. The efficiency comparison of the three methods above have been done on the previous paper[8].

These methods have good performance to handle missing values on the small size of data that has a few missing values and has less impact on the further data processing. The purpose of this work is to evaluate the effect of missing data presentation by implementing K-Nearest Neighbour imputation (KNNI) method. KNNI can be used to predict the two data types both discrete and continuous. KNNI does not require the formation of predictive models for each item which is experiencing missing data. K-Nearest Neighbour method applies KNNI algorithms which are commonly used in the classification process to handle missing values. KNNI method can find the value of closest neighborhood data that has missing value on attributes in the amount of k [9–13]

System/component data is retrieved from the univariate of maintenance historical data of secondary coolant system RSG GAS (PA), ie: (PA01/CP001, PA01/CR001, PA01/AH001, PA01/BT001, PA02/AP002, PA02/BT001). Matlab was used to build code of handling missing values. This code consists of three steps, such calculate distance, examine KNN method and calculate the efficiency of this method and compare to listwise deletion and mean substitution method with Root Mean Square Error (RMSE) using Matlab software.

The expected results of this study are to implement and obtain the best method to handle missing values for the evaluation of systems/components maintenance historical data of RSG GAS.

THEORY

Missing Value

Missing values are data or information that are missing or unavailable on the subject of research in a particular variable due to non-sampling error factors. Missing values have small impact on the final result when the number of missing values is small or small-sized data. However when number of missing values is very large then they greatly affect the final results of data analysis or decrease the accuracy. Missing values are classified into three categories: Missing Completely at Random (MCAR): The missing values have no relation on any other variable. Missing at Random (MAR): The missing values have relation to other variables. The missing values can be estimated by viewing the other variables. Meanwile, Missing Not at Random (MNAR): The missing values have relation to other missing values therefore missing data estimation cannot be performed by using the existing variables [2, 6].

Listwise Deletion

Listwise is the most common way to deal with missing data in research. Missing data is removed from the analysis. Recently, the data estimation techniques have been used to treat the missing values. The study shows if there are a lot of missing data then listwise deletion method will have a bias parameter with a large standard deviation, therefore accuracy of the estimation will reduce [2, 3, 6].

Mean Substitution

In order to avoid the decrease of sample number, thus if there is missing data in a research then imputation method will be conducted. Imputation is the process of replacing the missing value in the data set. Imputation method is the easiest way to replace missing values with mean value or mode. This method produces a mean estimation that is equal to the value imputed. The advantage of this method is to fill in missing values with the expected values. The expected values have relatively high stability values. The weakness of this method is the estimation error obtained by this method is always lower than the actual[2, 3, 6, 7].

KNN Imputation

K-Nearest Neighbour Imputation is a method that applies the nearest neighbor technique. This method is commonly used at imputation process or missing values fulfillment. It takes the values of the nearest neighbor to fill in the missing values which has the same attributes. The number of neighbors taken is the same depends on the entried *k* value, the number of the nearest observation.

KNNI can be used to predict the two data types both discrete and continuous data. KNNI does not require the formation of predictive model for each item which has missing values. The weakness of KNNI method is to find the lost value, KNN imputation algorithm will search through all the dataset. However, KNN imputation method is a method that is good enough for a missing value imputation[1, 4, 5, 9–11, 13].

The following steps are used for KNN imputation [4, 11, 13].

1. Determine the k value.
2. Calculate the distance between observation with missing data on variable-j and other observation without missing data by using Equation (1):

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{m} (x_{aj} - x_{bj})^2} \quad (1)$$

where:

$d(x_a, x_b)$ = distance between the target observation, x_a and observation, x_b.

x_{aj} = value of the variable-j on the target observation $xa_{.,}$ *J = 1, 2, 3, ..., m*

x_{bj} = value of the variable-j on the other observation x_b., j = 1, 2, 3, ...m

3. Find *k*, nearest observations, based on the value of the smallest distance. This value is used for imputation process on the observation with missing data.
4. Calculate the weight of each *k* of nearest observation.
5. Calculate the average of *k* of nearest observation without missing data by using the *weighted mean estimation* formula as follows:

$$\hat{x}_j = \frac{1}{W} \sum_{k=1}^{K} w_k v_{kj} \quad (2)$$

P-ISSN: 1411–240X E-ISSN: 2527-9963
Nomor: 632/AU3/P2MI-LIPI/03/2015

where v_{kj} are the values of the variable-j at observation k, k=1, 2, ..., K; $W = \sum_{k=1}^{K} w_k$, w_k are the nearest observation weights of k, which is formulated as follows: $w_k = \frac{1}{d(x, v_k)^2}$

6. Conduct imputation process of missing data on the observation with missing value with the obtained average value at step 5.

Evaluation criteria

Imputations method were compared based on tree measures of performance. Root mean square error (RMSE) measures the difference between imputed and true values and is the figure of merit employed by most studies. Basically, it represents the sample standard deviation of that difference [5, 6, 9] :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} x_i - \bar{x}}{n}} \quad (3)$$

where: xi = data-i, $\bar{x}$ = mean of data and n = number of sample size

METHODOLOGY

The purpose of this work is to evaluate the effect of missing data by using KNN method. Data was obtained from maintenance historical of RSG system/components of the secondary coolant system (PA), ie: (PA01/CP001, PA01/CR001, PA01/AH001, PA01/BT001, PA02/AP002, PA02/BT001). Evaluation of missing data effect were conducted with imputation methods, such as listwise deletion, means substitution and KNNI by using Matlab software. There are four steps to carry out the missing value imputation by using KNNI method, as follows:

1. Entry the data from attributes with missing values in excel format.
2. Calculate the distance to the nearest neighbors (k).
3. Examine KNN method to see how the number of nearest neighbors (*k*) during imputation. Examination is done by taking different k values to know the effect of changes in the results of imputation *k*.
4. Calculate the efficiency of KNN method and compare with other methods with Root Mean Square Error (RMSE) using Matlab software.

RESULTS AND DISCUSSION

Table 1 shows systems/components maintenance data of RSG GAS for secondary coolant pump (PA), on a terrace 53 to 88 (2005 to 2015). The missing values are denoted as NaN.

Table 1. Maintenance Data for 6 components of RSG GAS

NO	PA01/CP001 N=8	PA01/CR001 N=13	PA01/AH001 N=11	PA01/BT001 N=15	PA02/AP 002 N=6	PA02/ BT001 N=13
1	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	8	7
3	8	NaN	1	12	23	6
4	7	NaN	NaN	9	14	7
5	4	NaN	10	8	5	5
6	14	NaN	16	5	26	4
7	13	3	10	3		17

NO	PA01/CP001 N=8	PA01/CR001 N=13	PA01/AH001 N=11	PA01/BT001 N=15	PA02/AP 002 N=6	PA02/ BT001 N=13
8	8	8	2	8		12
9		7	21	1		7
10		3	6	3		3
11		7	24	7		3
12		12		11		8
13		7		9		8
14				6		
15				9		

Listwise deletion, mean substitution and KNNI were applied to handle missing values which are implemented at 6 maintenance historical data of primary coolant system/component.

Figure 1 shows the maintenance data in secondary coolant system (PA) on the six components with missing data. The missing data on the PA system for six components, namely PA01/CP001, PA01/CR001, PA01/AH001, PA01/BT001, PA02/AP002, and PA02/BT001, are 2, 6, 3, 2, 1, and 1, respectively. Numbers of data from those six components are 8, 13, 11, 15, 6, and 13, respectively. The percentage of missing data of those six components are 25%, 46%, 27%, 13%, 17% and 8%, respectively.

Listwise deletion method does not include the missing data in calculation. The mean value is obtained from the existing data. The results of the imputation to those six components are: 9, 6.71, 11,25, 7,00, 15.2 and 7.25, respectively.

	A	B	C	D	E	F	G
1	No	PA01/CP001	PA01/CR001	PA01/AH001	PA01/BT001	PA02/AP002	PA02/BT001
2	1						
3	2					8	7
4	3	8		1	12	23	6
5	4	7			9	14	7
6	5	4		10	8	5	5
7	6	14		16	5	26	4
8	7	13	3	10	3		17
9	8	8	8	2	8		12
10	9		7	21	1		7
11	10		3	6	3		3
12	11		7	24	7		3
13	12		12		11		8
14	13		7		9		8
15	14				6		
16	15				9		

Figure 1. Results of Listwise Deletion Method Implementation

Figure 2 shows the results of imputation using mean substitution method. Data from the 6 columns using listwise deletion method values were considered as mean values. Then mean values of the column (1) to (6) will be filled into the missing data. Finally, the previous steps generated the complete data in columns (1) to (6), and furthermore produced the mean values. The mean values were generated using mean substitution method that equal to the mean values in the listwise deletion method.

	A	B	C	D	E	F	G
1	No	PA01/CP001	PA01/CR001	PA01/AH001	PA01/BT001	PA02/AP002	PA02/BT001
2	1	9	6.71	11.25	7	15.2	7.25
3	2	9	6.71	11.25	7	8	7
4	3	8	6.71	1	12	23	6
5	4	7	6.71	11.25	9	14	7
6	5	4	6.71	10	8	5	5
7	6	14	6.71	16	5	26	4
8	7	13	3	10	3		17
9	8	8	8	2	8		12
10	9		7	21	1		7
11	10		3	6	3		3
12	11		7	24	7		3
13	12		12		11		8
14	13		7		9		8
15	14				6		
16	15				9		

Figure 2. Results of Mean Substitution Method Implementation

P-ISSN: 1411–240X E-ISSN: 2527-9963
Nomor: 632/AU3/P2MI-LIPI/03/2015

Figure 3 shows the results of KNNI algorithm calculation employing equations (1) and (2) for k = 1 to n for PA02/BT001. The estimated X value was imputed to the missing data in the data set. RMSE values were then determined for each value of k.

From the calculation results for PA02 /BT001, it was estimated that X = 7.068966 for k = 5. This X estimated value was imputed into the missing data. Mean and Root Mean Square Error (RMSE) was eventually determined. Mean and RMSE values are 7.236074 and 3.640375.

	A	B	C	D	E	F	G	H	I	J	K	L
1	No	X	(Xa-Xb)2	Sqrt(Xa-Xb)	d	SortX	K	Wk	W	Xest	Mean	RMSE
2												
3	1	7	0.0625	0.0625	0.25	7	1	16	16	7	7.230769	3.640664
4	2	6	1.5625	0.0625	0.25	7	2	16	32	7	7.230769	3.640664
5	3	7	0.0625	0.0625	0.25	7	3	16	48	7	7.230769	3.640664
6	4	5	5.0625	0.5625	0.75	8	4	1.777778	49.77778	7.035714	7.233516	3.640503
7	5	4	10.5625	0.5625	0.75	8	5	1.777778	51.55556	7.068966	7.236074	3.640375
8	6	17	95.0625	1.5625	1.25	6	6	0.64	52.19556	7.055858	7.235066	3.640423
9	7	12	22.5625	5.0625	2.25	5	7	0.197531	52.39309	7.048107	7.23447	3.640452
10	8	7	0.0625	10.5625	3.25	4	8	0.094675	52.48776	7.042609	7.234047	3.640474
11	9	3	18.0625	18.0625	4.25	3	9	0.055363	52.54321	7.03835	7.233719	3.640492
12	10	3	18.0625	18.0625	4.25	3	10	0.055363	52.59849	7.034099	7.233392	3.64051
13	11	8	0.5625	22.5625	4.75	12	11	0.044321	52.64281	7.03828	7.233714	3.640492
14	12	8	0.5625	95.0625	9.75	17	12	0.010519	52.65333	7.04027	7.233867	3.640484

Figure 3. Results of KNNI Method on PA02/BT001

From Figure 3 RMSE values for each k = 1 to 12 were obtained. The smallest RMSE value is obtained when the amount of k = 5. Comparison of RMSE values are shown in Figure 4.

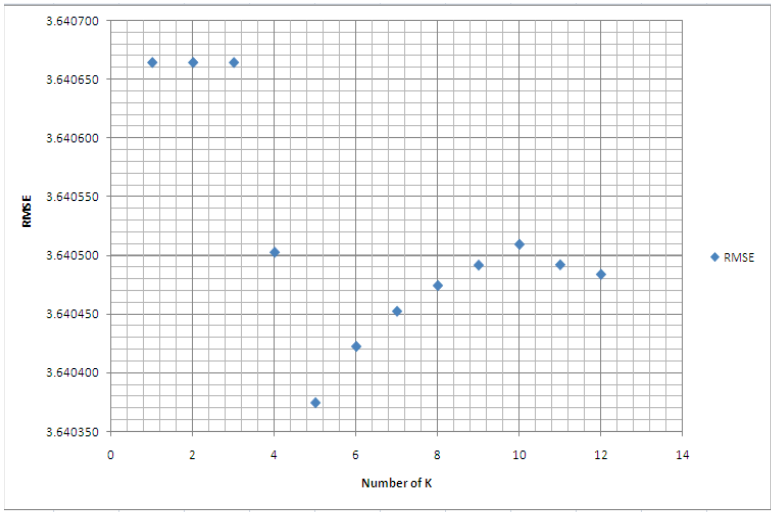

Figure 4. Variation of RMSE Value with Number of k from KNNI Method Implementation

Table 2 represents the values of mean and Root Mean Square Error of the missing values obtained from the implementation of listwise deletion, substitution means, and KNNI methods. Root Mean Square Error for listwise deletion and mean substitution method were obtained using Equation (3), where mean values are based on Figure 1-3. The number of data (n) for mean and RMSE calculation in the deletion listwise method for six components are 6, 7, 8, 13, 5, and 12. While the number of data in the mean substitution and KNNI method for all six components are 8, 13, 11, 15, 6 and 13.

Mean and RMSE values for PA02/BT001 using listwise deletion, substitution means, and KNNI method are 7.2500, 7.2500, 7.2361 respectively. These results indicate that there is similarity for listwise deletion and mean substitution method, but differ with KNNI. It is due to imputation is conducted based on the mean values of the existing data and mean of KNN imputation method is determined based on the minimum distance. RMSE values for these three methods are 3.95716, 3.78869, and 3.640375. From these three methods, KNNI method produces the smallest RMSE

values. RMSE comparison chart of these methods are shown in Figure 5. The mean values of these three methods are shown in Figure 6.

Table 2. Means dan RMSE of Handling With Missing Value Method

System/ Component	Liswise Deletion		Mean Substitusi		K–Nearest Neighbour Imputation (KNNI)	
	Mean	RMSE	Mean	RMSE	Mean	RMSE
PA01/CP001	9.000	3.79473	9.000	3.20713	8.75	3.02751
PA01/CR001	6.7143	3.09377	6.7143	2.18763	6.6462	2.10663
PA01/AH01	11.250	8.46421	11.250	6.75210	10.964	6.46825
PA01/BT001	7.0000	3.26599	7.0000	3.02372	6.7321	2.83425
PA02/AP002	15.200	9.14877	15.200	8.18291	15.011	7.481908
PA02/BT001	7.2500	3.95716	7.2500	3.78869	7.2361	3.640375

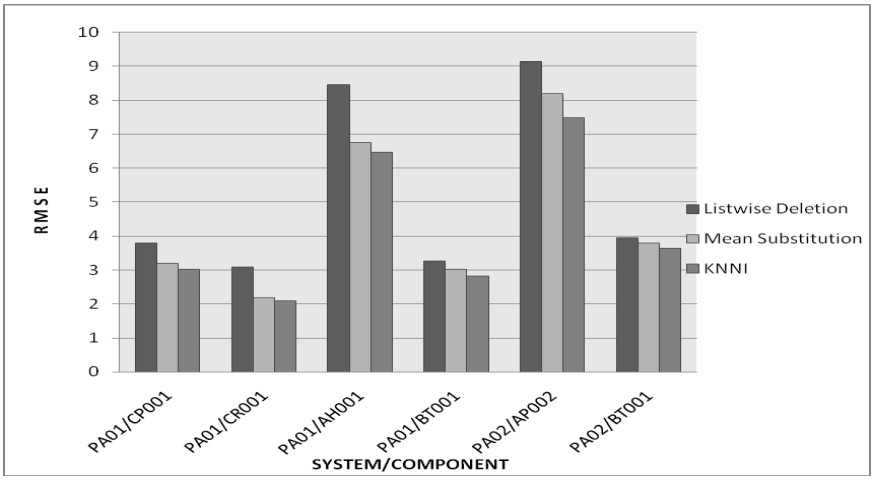

Figure 5. Variation of RMSE Values on Systems/Components using Different Methods

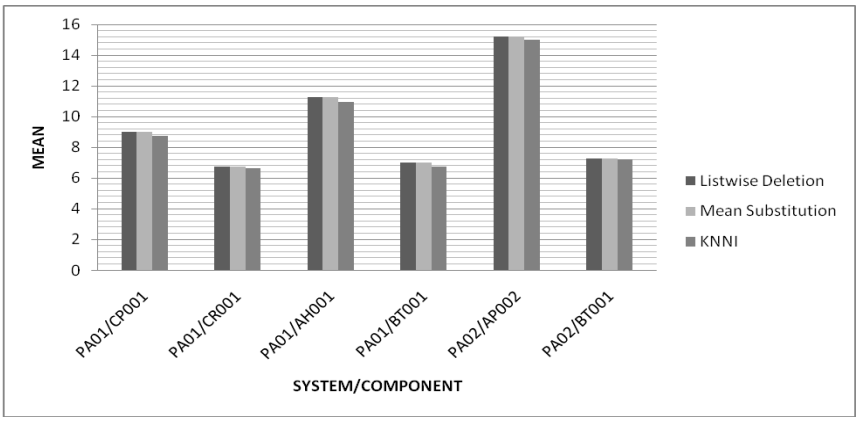

Figure 6. Variation of Mean Values on Systems/Components using Different Methods

The missing data on maintenance historical data of secondary coolant system (PA) for the six components, namely PA01/CP001, PA01/CR001, PA01/AH01, PA01/BT001, PA02/AP002, and PA02/BT001 can be replaced with mean values of 8.75, 6.65, 10.96, 6.73, 15.01 and 7.24, respectively which are generated by KNNI method.

P-ISSN: 1411–240X E-ISSN: 2527-9963
Nomor: 632/AU3/P2MI-LIPI/03/2015

CONCLUSION

The missing values handling evaluation in the secondary coolant system (PA) RSG GAS on terrace 53 to 88 (2005 to 2015) ie: PA01/CP001, PA01/CR001, PA01/AH001, PA01/BT001, PA02/AP002, and PA02/BT001 were carried out. It can be concluded that imputation method based on Machine Learning K-nearst Neighbor imputation (KNNI) is the best way as compared to listwise deletion and Mean Substitution imputation methods. It produces the smallest RMSE value. Meanwhile the trend of the RMSE of the three methods from the smallest can be sequenced as follows: listwise deletion, mean substitution and K-Nearest Neighbour imputation.

ACKNOWLEDGEMENT

Author would like to thank to Syaiful Bakhri, Ph.D, Drs. Deswandri, M.Eng and Aep Saepudin C, ST. who had given advices and guidances during this research project. The research has been funded by BATAN DIPA 2016.

REFERENCES

1. Umathe, Vaishali H G.C. A Review on Incomplete Data And Clustering. Int. J. Comput. Sci. Inf. Technol. 2015. **6**(2):1225–7.
2. Dong Y., Peng C.J. Principled missing data methods for researchers. Springer Plus. 2013. **2004**:1–17.
3. Arumuga Nainar S A Comparative Study of Missing Value Imputation Methods on Time Series Data A Comparative Study of Missing Value Imputation Methods on Time Series Data. Int. J. Tecnol. Innov. Res. 2015. **14**:1–8.
4. Minakshi, Rajan Vohra G. Missing Value Imputation in Multi Attribute Data Set. Int. J. Comput. Sci. Inf. Tecnol. 2014. **5**(4):5315–21.
5. Nookhong J., Kaewrattanapat N. Efficiency Comparison of Data Mining Techniques for Missing-Value Imputation. J. Ind. Intellegent Inf. 2015. **3**(4):305–9.
6. Schmitt P., Mandel J., Guedj M. Biometrics & Biostatistics A Comparison of Six Methods for Missing Data Imputation. J Biomet Biostat. 2015. **6**(1):1–6.
7. Somasundaram R.S., Nedunchezhian R. Missing Value Imputation using Refined Mean Substitution. Int. J. Comput. Sci. 2012. **9**(4):306–13.
8. Hartini E. Efficiency Comparison of Method of Handling Missing Value In Data Evaluation System or Component. 2016.
9. Arslan I.B.A. and A. A Novel Hybrid Approach to Estimating Missing Valuse in Databases Using K-Nearest Neighbors and Neural Networks. Int. J. Innov. Comput. Inf. Control. 2012. **8**(7):4705–17.
10. Malarvizhi M.R., Thanamani A.S. K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation. J. Comput. Eng. 2012. **6**(5):12–5.
11. Malarvizhi M.R., Thanamani A.S. K-Nearest Neighbor in Missing Data Imputation. Int. J. Eng. Res. Dev. 2012. **5**(1):5–7.
12. Suguna N., Thanushkodi K.G. Predicting Missing Attribute Values Using k-Means Clustering. J. Comput. Sci. 2011. **7**(2):216–24.
13. Wael M.Khedr A.M.E. Pattern Classification for Incomplete Data Using PPCA and KNN. J. Eng. Trends Comput. Inf. Sci. 2013. **4**(8):628–32.