Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 1, January 2020, pp. 543~550
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17.i1.pp543-550

Proposed algorithm for Regression-based prediction with bulk noise

Chanintorn Jittawiriyanukoon
Graduate School of Advanced Technology Management, Assumption University, Thailand

Article Info

Article history:

Received Apr 22, 2019
Revised Jun 23, 2019
Accepted Jul 1, 2019

Keywords:

Bulk noise
Mean absolute error
Mean square error
Noisy and missing data
Regression-based prediction
Signal-to-noise ratio

ABSTRACT

The noise has incited an original data due to a network with an inferior SNR. In case of the bulk noise, the insightful content within the data is substantially squeezed. A cost-effective method will challenge to quarantine the insights, so that information can be utilized more resourcefully. To achieve this aim, it is essential to iron the bulk noise content out, and then calculate the analytics of the clean data. As noise is bulk so some statistical methodologies such as averaging or randomizing are employed. A prediction using the regression-based model with bulk noise for the experiment in practice is introduced. The decomposition approach to separate the insights is exploited. The proposed algorithm achieves a (local) solution at each computing step and selects the best solution in view of global impacts. The correlation coefficient, average error, absolute error and mean squared error are used to constitute the prediction. Results from MOA simulation will be compared to actual data in the succeeding time. The prediction with bulk noise using the proposed algorithm outperforms other imputation methods.

Corresponding Author:

Chanintorn Jittawiriyanukoon,
Graduate School of Advanced Technology Management,
Assumption University,
Samut Prakan Province, 10540, Thailand.
Email: pct2526@yahoo.com

1. INTRODUCTION

The missing data are pervasive in the calculating practice. They can be missing for some instances or attributes [1-3]. If mainstream (bulk) of data is missing on the attribute then it is alleged to be unnoticed. Traditional treatments and software always assume that all attributes in a dataset are figured for all instances. The popular method for all fundamental software is to eliminate instances with any noise a technique is known as complete data analytics [4-6]. The evident weakness of elimation is that in case of bulk noise, it habitually cancels a hefty portion of the attribute, resulting to a bold loss of numerical implication. Data scientist is plausibly unwilling to abandon data he has spent money, effort, and time in accumulating. As such, most treatment techniques for the case with bulk noise have become prominent.

Pampaka, et al [7] define missing values as the noise which is not deposited for an entity in the instance of interest. The complication of missing value is corporate in most researches and reflects nontrivial conclusions. Many types of research have attended to treat the noise and problems arisen from missing values, and the approaches to prevent particularly in the medical area [8]. Dziura, et al [9] introduce the promising approach of treating the noise is to avoid the issue by well-design the study and amassing the data prudently. Mallinckrodt, et al [10] are signifying to lessen the amount of noise in the scientific study. They propose the planning has to edge the data accumulation to researchers. This can be attained by decreasing the number of critical data collecting, investigations, and using the befitting visualization. Prior to the study, a comprehensive documentation of the research is to prepare the guide of operations including the ways to select the members, procedure to train the members, the noise treatment, as well as process to collect and

revise the data. Besides, if a trivial project is targeted before the primary collection, it may help detect the unpredictable complications which may arise during the research, as well as sinking the number of missing values.

In repetition, bulk noise [11] cultivates whenever unrecognized characters including null, blank, and others have occupied any rows as shown in the table of Figure 1. Noise data can cultivate an erratic consequence varying from the erroneous dataset to nonresponsive execution. Virmani, et al [12] introduce a clustering algorithm based on K-means in order to rally results for users over social networks. The K-means algorithm per se allows the researcher to fix the K value. The paper based on the fixed figure of K improves 70% in similarity experiment. Shi, et al [13] investigate an innovative algorithm to opt the fitness calculation to the union function in K-Means algorithm. Results based upon the combination of these functions afford a better comprehensive document. Wartana, et al [14] introduce a Fuzzy-based algorithm to increase the security and stability of the power system. It proves that the fuzzy algorithm is supporting the decision making more effectively than the genetic algorithm. Manoj et al [15] propose the predictive framework based on the neural network model for optimal performance of the reusability of the code. The least square algorithm also is used to obtain optimization in order to calculate and confirm the highest reliability.

Bulk noise represents any unreadable and useless data which is collected unintentionally, but obscures. Suresh et al [16] treat a denoised process to improve the spectral of satellite image. These Gaussian noises are contaminating not only corrupted problems such as hardware or software incompatibility but also processing vulnerabilities such as no further execution, or no operation, or failure. A bulk noise can ruin the classifying process of the dataset. In this case, bulk noise worsens the stability analysis and remains an excessive risk. To denoise satellite images is critical for improving the visualization of images and for easing supplementary analysis and its processing tasks.

	Total Amount				
	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5
Trans A	10234	XX		*&^	
Trans B					
Trans C	234.6	CH	9076		!!!
Trans D			AZX		
Trans E	342.46			@#	N/A

Figure 1. An Example of Noise Pattern. Blank Indicates that the Value is Missing

The objective of the research is to investigate the accuracy of the regression model for bulk noise data using MOA [17]. In the analysis, a large portion of noise is found to be above fifty percent of the total size of the dataset. This is called, "bulk noise" which is illogical fluctuation due to attribute which is not able to be accounted for. Bulk noise will be considered from practical points of view. The noise part thus needs to be detected in order to break through the failure in manipulation. Next, the proposed algorithm will treat these noises then prediction results from simulation are collected to legalize the accuracy. Finally, the correctness of the proposed treatment will be compared with the actual data.

2. RELATED WORK

Conservative statistical computation and software count on collected instances in an indicated framework for entire cases. For a lengthy time, the missing data is explained as the ‘unknown’ of computation. Although most cases experience missing value and require treating the problem in some techniques, there is absolutely nothing found in the literature or practical guidance. It is so far because none of the widely used methods have any concrete calculations. A method for dealing with the missing values is presented [18] as the temporal data is unsurprisingly recurring using different discretization techniques. The concept of exclusion or inclusion of: a temporal sequence of the data, classification label, and managing of stream data for temporal data discretization is applied. The prerequisite is that data needs to persist. The authors [19] present the regression models where the primary relationship embraces interaction expressions. A linear framework with one fully witnessed predictor is considered. Then the conditional distribution of interaction expression and the missing covariance is applied for examining the performance of multiple imputations. Other techniques which can be employed by adjusting multiple imputation software to outperform in spite of incompatibilities between underlying relationships among the attributes and framework assumptions are investigated.

Nonetheless, the experiment in this research does not shadow any approaches as mentioned earlier. The proposed treatment begins with the unwanted bulk noise classification. After that, the proposed algorithm repairs all unwanted elements in the dataset by obtaining a local optimal solution at each computing step and chooses the best solution in view of global impacts. Note that even if the single element of noise in the dataset can impede the data running unless the exclusion of the noise. The existing two algorithms, namely, Mean Variables (MV), and Random Imputation (RI) are applied for repairing noise with substitution. Thus, the computation costs which are inclusive of searching time for bulk noise removal and algorithm run time will be cited. These two algorithms are compared with actual values to reflect their precisions. The experimental results using MOA simulation are collected to check the accuracy between existing and proposed algorithms. The awaiting outline of the research is as follows. In section III, bulk noise conditions are introduced. Section IV explains the performance results of the proposed algorithm from experimental perspective. Section V finally outlines the conclusion of the research.

3. BULK NOISE CHARACTERISTICS

Characteristics of bulk missing values are discussed, datasets with bulk noise are illustrated in this section. Note that a few entries of noise can crook a dataset as the whole. Bulk noise can develop much higher impact than ever as it can certainly create faults during data compiling or storing. A noise blocks the insight extraction in data curation, which can result in the aborted deep learning operation. It can be frantically complex to leverage the faults. As such, to classify and treat the noise data are a must to overcome the constraint. In this research, the overwhelm case of noise in the dataset is studied. Bulk noise revenues the attendance of noise in the dataset to be outside 50%. The convolution is to quest systematically where the bulk noise accompanies. The search concludes the essence of the bid of noise treatment. To terminate bulk noise, the deterministic dataset at hand for execution is assumed. In this research, a split-and-repair is taken on by expecting that a dataset D can be split into two parts: a minor but clean part, Dc and a bulk noise part, Dn. In the noisy environment ($Dn \geq Dc$), the assumption is more representative. However, in case of the gigantic dataset, to purge bulk noise is ascending up the split-and-repair time correspondingly. The simulation on the dataset with bulk noise displays the sufficient performance accordingly.

A general approach to deal with bulk noise data is to purge all instances containing the noise. But, the technique as such will not iron out the bulk noise problem as, only a Dc remains. Not to mention, removed instances can affect the ongoing data curation. To screen Dn in the dataset, the existent bound of the noise is presumed. Then, optimization is probable on the simulation.

The split-and-repair method for Dn is a main target of the research as bulk noise unless purging can discontinue further data analytics. Two approaches for estimating data for Dn which are Mean Variables (MV), and Random Imputation (RI) have been introduced. Let D be a dataset matrix which contains a rows and b columns, while n represents instances affected by noise, in which n is always less than a ($n < a$ and Dn1, Dn2, Dn3,…, Dn(b-1), Dnb) for each n = 1, 2, 3,…, a. The D matrix is expected to be a deterministic set. An element Dnb is set of the noisy element whenever $\{Dij = \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$. Remark that in case of bulk noise, $n \geq a/2$. The dataset with bulk noise is called troubled dataset. Hence, the proposed treatment to revolve the hazard and continue the analysis by applying the estimated vector En is described in the next section.

The split-and-repair strikes out noise which can be screened by an impaired filtering, but eliminated instances can hamper the analytics. Noise can misinterpret to negative, inducing data science to keep on with fault decision (a type one error). In order to assure data analysis, these Dkb must be definitely denoised. It is crucial to detach Dn, particularly for the bulk noise where $n \geq a/2$, any techniques have to stress on a remaining minor fraction of the whole dataset. This research motivates the proposed algorithm for bulk noise. The simulation is based on the regressive model with ten synthetic datasets. In the individual experiment, the simulation is run for the proposed algorithm, Mean Variables (MV) and Random Imputation (RI) after denoising. The results from three treatments will be compared to those actual data in the subsequent year.

4. RESULTS AND ANALYSIS

The MOA simulation is designated for analyzing ten datasets. The investigation of a regression model for bulk noise level (n) is performed. The study is deployed on an Intel® Core ™ i5 CPU, 1.60 GHz Processor and 8 GB RAM on board. The datasets are diverse in file size, instances, and attributes.

4.1. Correlation Coefficient (COEF)

The COEF is one of the metrics in the statistics. It is a useful analysis which calculates the power concerning connections and variables. In statistics, this coefficient refers as the R-test. It defines how

powerful connection among two variables is. The figure ranges between 1.0 and -1.0. If the figure is negative then, it determines if one declines, the other rises. Also, if the figure is positive, then it earns both of them either lessen or grow collectively. The computation for this metric can be found in [20].

4.2. Mean Square Error

Mean squared error (MSE) [21] is one of many types in statistics to enumerate the differences among the sample and population awaited by a regression model. The lower the MSE, the nearer to the best-fit curve is concluded. The MSE clarifies the standard statistical metric of the dissimilarity among observation and forecast. The different figure is calculated by the targeted data over the error in the forecast. A dataset in a working set drops the error value for the experiment dataset. Fault rate for training dataset will be comparatively higher than that of the experiment set. If any two algorithms produce the like mean absolute error then MSE is deployed for a decision, which is the optimum answer.

4.3. Mean Absolute Error

The mean absolute error (MAE) [21] is a figure deployed to evaluate the fussy forecasts. The MAE is an average of the absolute figure of faults and can be defined as model evaluation statistics.

4.4. Mean Variables (MV)

Mean value criterion [22] is to assign data for all n instances. Apply the split-and-repair to the D dataset and classify Dn, a dataset comprises of n instances with noise. Any n rows of the matrix D possess an element dij with noise data where {dij = ϕ || ∞, 1 ≤ i ≤ n; 1 ≤ j ≤ b} then the row is swapped by the MV for estimated En dataset as listed in (1):

$$d_{ij} = \frac{1}{|a-n|} \sum_{x=n+1}^{a} d_{xj} \quad (1)$$

The investigation of the MV is that it is an acceptable forecast for a parameter out of a normal distribution. This treatment somehow induces a volatile unfairness. Not to mention the MV is led by the slanted replacement as well as cultivates the size of state space.

4.5. Random Imputation (RI)

Utilize several imputations at random for replacement. Analogous to the above MV, the split-and-repair is applied to the targeted D dataset and results a dataset with n instances. Any n rows of the matrix D possess an element dij with noise data where {dij = ϕ || ∞, 1 ≤ i ≤ n; 1 ≤ j ≤ b} then the row is switched by the RI for estimated En dataset. The minimum likelihood found in column j (where j = 1, 2, 3,…,b) is marked by d(min)j where d(min)j = Min (dnj) for each n = 1, 2, 3,…, (a-n). Likewise, the maximum likelihood of column j (where j = 1, 2, 3,…,b) is defined by d(max)j where d(max)j = Max (dnj) for each n = 1, 2, 3,…, (a-n). The substitution for estimated En dataset with multiple imputations for n instances in each column j is randomly explained as follows:

$$d_{ij} = RAND\left[d(min)_j, d(max)_j\right] \quad (2)$$

4.6. Proposed Algorithm

The proposed algorithm works straightforwardly, as described in the following stages. The dataset will be split into Dc and Dn. The Dc portion is assumed to provide the solution. In general, it is the split-and-repair approach. The successful calculation to cover up Dn in every fractional step imposes on the fruitful calculation of every subsolution. This is called the optimal features as an optimal solution can be made out of optimal subsolutions. To reach accomplishment at each partial step, the proposed algorithm contemplates the subsolution data only at that partial step. Namely, the decision of each fractional step the proposed algorithm makes is based on a global consequence. This will complete a global policy to obtain the optimal characteristic and is sufficient to compromise decisive goal. As a metaphor, it's analogous to doing the chess by keeping thinking ahead more than one move, and finally scoring the game. The proposed algorithm needs no complex decision rule as it only deliberates all the available subsolutions at each stage. There is not necessary to calculate feasible decision inferences then the computation cost is about O(ab). The proposed algorithm is summarized in Figure 2.

State space is nontrivial to reflect the speed of computing complexity. In this research, the computation cost is derived, corresponding to the performance assessment. It is deceptive any forecasts are problematical if the computation cost is extraordinary as depicted in Table 1. Note that in case of bulk noise, a is always smaller than 2n.

Proposed Algorithm

Require: Data matrix $[D]_{xy}$ with x rows and y columns
Ensure:$[D]_{xy}$, S = all potential solutions in each computation step = $\{S_1^t, \ldots, S_y^t\}$, C_y^t = centroid of the attribute y, O_y^t = candidates in each computation step, P_r = a premium solution where $P_r(S_y) \geq 0$ and $S_y \in S^t$
for $I = 1$ to x do
for $J = 1$ to y do

$O_y^t \leftarrow 0$
for $k = 1$ to S /** All solutions computation **/
$F_k = \arg\max_{S_y \in S^t} P_r(S_k)$
/** Solution F_k and corresponding C_k **/
end for
for $n = 1$ to S /** Choose best solution **/
$\Delta_n = |F_n - C_n|$
end for
$O_y^t = \arg\min_{S_y \in S^t} (\Delta_1, \Delta_2, \Delta_3, \ldots, \Delta_s)$ /** A new best for this computation step **/
Return O_y^t /** Regression-based computation **/

end for
end for

Figure 2. Proposed algorithm

Table 1. Computation Complexity of Proposed Method

Treatment	Computation Complexity
MV	O(*ab*) + O(*ab*-*bn*) ≈ O(*ab*)
RI	O(*ab*) + O(2(*ab*-*bn*)) ≈ O(*ab*)
PROPOSED	O(*ab*) + O(2(*ab*-*bn*)) ≈ O(*ab*)

In this research, the split-and-repair strategy is proposed in order to handle the bulk noise. The strategy will split and repair the bulk noise portion prior to the forecast. Another model-based strategy will rather review the algorithm per se to leverage the noise before the use of the parametric forecast. The latter strategy can be found in either ANCOVA [23, 24] or PSPP application, which relates countless imputations for interchanging the noise. While the split-and-repair technique [25] gears prospect data to consideration. The model-based algorithm is somehow complex, and the user's skill is obligatory as it has been profoundly designed to replicate the parametric one. The error values of ten divergent datasets using MOA at noise value ranging from 50% to 80% are examined. This is a primitive analytics toward the nominated datasets, and all results are shown in Table 2-5. The three errors in the table distinguish the correlation coefficient (COEF), the mean squared error (MSE), and mean absolute error (MAE) individually. Dataset#2 gives lowest figure for COEF, MSE and MAE. The regression-based forecast is depicted in Table 6.

Table 2. Forecast with Mean Absolute Error for Ten Different Datasets (N = 0.5)

Dataset	COEF	MSE	MAE
1	0.31	17.2	14.2
2	0.08	1.83	1.61
3	0.29	28.7	24.9
4	0.17	30.2	26.3
5	0.28	67.4	57.3
6	0.79	3.08	2.3
7	0.14	49.1	37.9
8	0.32	12.7	10.2
9	0.04	15.7	13.6
10	0.2	20.1	14

Table 3. Forecast with Mean Absolute Error for Ten Different Datasets (N = 0.6)

Dataset	COEF	MSE	MAE
1	0.2	17	14
2	0.01	1.83	1.54
3	0.34	28.7	25.2
4	0.16	26.3	24
5	0.30	71.2	60.9
6	0.79	3.07	2.29
7	0	48.6	37.8
8	0.35	12.5	10
9	0.15	15.6	13
10	0.24	20	14.3

Table 4. Forecast with Mean Absolute Error for Ten Different Datasets (N = 0.7)

Dataset	COEF	MSE	MAE
1	0.3	17	14.1
2	0.18	1.81	1.59
3	0.34	28.7	25.2
4	28.8	0.28	25
5	0.19	71.2	61.5
6	0.79	3.08	2.29
7	0.01	48.9	38.5
8	0.35	12.5	10
9	0.07	14.9	12.9
10	0.21	20.1	14

Table 5. Forecast with Mean Absolute Error for Ten Different Datasets (N = 0.8)

Dataset	COEF	MSE	MAE
1	0.37	17.35	14.58
2	0.18	1.82	1.59
3	0.34	28.7	25.2
4	0.17	30.2	26.3
5	0.02	69.3	59.4
6	0.8	3	2.23
7	0.32	49.5	37.9
8	0.35	12.56	10
9	0.31	14.5	12.6
10	0.21	20.1	14

Table 6. Regression-Based Forecast for Ten Datasets

		Regression-based Forecast
DATASET	1	$X_9=0.838X_3+24.56$
	2	$X_{12}= -0.117X_2+3.89$
	3	$X_4=1.78X_7+147$
	4	$X_1=6.34X_3-6.2X_5-50.3$
	5	$X_6=0.28X_2+0.23X_3+1284.3$
	6	$X_7= 0.36X_4+0.18X_5+0.21X_6-18.09$
	7	$X_1=5.3X_4+0.23X_5+110.6$
	8	$X_3= -82.7X_2+0.07X_5+422.7$
	9	$X_6=-1.27X_2+624.53$
	10	$X_7= -1.4X_1+1.3X_3-1.4X_5-0.4X_4-18.5$

Tables 7-10 disclose an average error for the regression-based model associating to the authentic data. In this research, ten dissimilar datasets are studied at the divergent noise level (n) is extending from 50% to 80% as charted in Table 7-10 correspondingly. In very cases of the forecast from the proposed method, the error is lowest. Moreover, in case of bulk noise, the computation complexity for all three treatments is akin. It concludes the proposed method is the utmost effective algorithm for bulk noise analytics.

Table 7. Average Percentage of Error for Ten Different Datasets (N=0.5)

	n = 0.5		
Dataset	MV	RI	PROPOSED
1	14.15	14.16	13.67
2	16.03	16.22	15.83
3	17.6	16.9	15.66
4	35.27	36.5	24.7
5	32.5	35.4	7.4
6	10.23	18.14	9.9
7	51.4	58.5	41.3
8	13	13.9	11.8
9	54.5	62	46
10	17.71	17.23	15.23

Table 8. Average Percentage of Error for Ten Different Datasets (N=0.6)

n = 0.6			
Dataset	MV	RI	PROPOSED
1	14.08	14.21	13.65
2	16.08	16.28	15.83
3	17.6	16.9	15.66
4	35.2	36.5	24.2
5	28.3	30.4	6.4
6	10.95	23.15	10.1
7	51.2	49.9	37.5
8	13.1	13.6	11.8
9	54.9	60.8	47.1
10	20.76	20.73	18.29

Table 9. Average Percentage of Error for Ten Different Datasets (N=0.7)

n = 0.7			
Dataset	MV	RI	PROPOSED
1	14.22	14.06	13.52
2	16.08	16.2	15.72
3	17.6	16.9	15.66
4	35.2	36.5	24
5	27.4	29.6	5.32
6	12.07	29.4	11.1
7	51	51.1	39.3
8	13.1	14.3	11.9
9	52.3	53.1	41.6
10	23.5	21.39	18.95

Table 10. Average Percentage of Error for Ten Different Datasets (N=0.8)

n = 0.8			
Dataset	MV	RI	PROPOSED
1	13.98	14.69	13.44
2	16.08	15.97	15.62
3	17.6	16.9	15.66
4	35.2	36.5	24
5	27.4	28.9	4.52
6	11.9	30.2	11.6
7	51.4	49.3	38
8	13.1	13.3	11.8
9	55.5	54.4	44.2
10	21.06	18.5	17.6

5. CONCLUSION

In this paper, conventional algorithms for treating noise are imperfect. Under the certain condition, they seriously harvest both standard error and biased parametric forecast. Not to mention, the conservative imputations, MV and RI mechanisms, yield severe average error figures. The proposed mechanism is proven to be a benign choice when forecasting regression models for which optimum solution is concerned. It also exhibits the benefit of not demanding the extra computation cost. Next move will investigate other different imputations, so that the suitable suboptimal solution in each computation phase will be further investigated.

REFERENCES

[1] X. Zhu, S. Zhang, Z. Jin, and Z. Zhang, "Missing Value Estimation for Mixed-Attribute Data Sets", *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110-121, 2011.
[2] T. Deng, W. Fan, and F. Geerts, "Capturing Missing Tuples and Missing Values", *ACM Transactions on Database Systems*, vol. 41, issue. 2, pp.10:1-10:47, 2016.
[3] M. M. Rahman and D. N. Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets", *IAENG Transactions on Engineering Technologies*, pp. 245-257, 2013.
[4] V. Boeva, L. Lundberg, M. Angelova, and J. Kohstall, "Cluster Validation Measures for Label Noise Filtering", *International Conference on Intelligent Systems*, pp. 109-116, 2018.
[5] B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: a Survey", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845-869, 2014.
[6] M. Basner, *et al.*, "Auditory and non-auditory effects of noise on health", *Lancet*, pp. 1325–1332, 2014.
[7] M. Pampaka, G. Hutcheson and J. Williams, "Handling missing data: analysis of a challenging data set using multiple imputation", *International Journal of Research and Method in Education*, vol. 39, no. 1, pp. 19-37, 2016.

[8] R. T. O'Neill and R. Temple, "The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It", *Clinical Pharmacology and Therapeutics*, vol. 91, no. 3, pp. 550-554, 2012.
[9] J. D. Dziura, L. A. Post, Q. Zhao, Z. Fu, and P. Peduzzi, "Strategies for dealing with Missing data in clinical trials: From design to Analysis", *Yale Journal of Biology and Medicine*, vol. 86, pp. 343-358, 2013.
[10] C. Mallinckrodt, et al, "Recent Developments in the Prevention and Treatment of Missing Data", *Therapeutic Innovation and Regulatory Science*, vol. 48, no. 1, pp. 68-80, 2013.
[11] C. Enders, "Applied Missing Data Analysis", *Guilford Press*, New York, 2010.
[12] C. Virmani, A. Pillai and D. Juneja, "Clustering in Aggregated User Profiles across Multiple Social Networks", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 6, pp. 3692-3699, 2017.
[13] K. Shi and L. Li, "High performance genetic algorithm based text clustering using parts of speech and outlier elimination", *Applied Intelligence*, vol. 38, pp. 511-519, 2013.
[14] I. M. Wartana, N. P. Agustini and J. G. Singh, "Optimal Integration of the Renewable Energy to the Grid by Considering Small Signal Stability Constraint", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2329-2337, 2017.
[15] H. M. Manoj and A. N. Nandakumar, "A Novel Optimization towards Higher Reliability in Predictive Modeling towards Code Reusability", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2855-2862, 2017.
[16] S. Suresh, S. Lai, C. Chen and T. Celik, "Multispectral Satellite Image Denoising via Adaptive Cuckoo Search-Based Wiener Filter ", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no.8, pp. 4334-4345, 2018.
[17] A. Bifet, R. Kirkby, G. Holmes and B. Pfahringer, "MOA: Massive Online Analysis", *Journal of Machine Learning Research*, vol. 11, pp.1601-1604, 2010.
[18] P. Chaudhari, D. P. Rana, R. G. Mehta, N. J. Mistry and M. M. Raghuwanshi, "Discretization of Temporal Data: A Survey", *International Journal of Computer Science and Information Security*, vol. 12, no. 2, pp. 66-69, 2014.
[19] S. Kim, C. A. Sugar and T. R.Belin, "Evaluating model based imputation methods for missing covariates in regression models with interactions", *Statistics in Medicine*, vol. 34, no. 11, pp. 1876-1888, 2015.
[20] M. M. Mukaka, "A Guide to Appropriate Use of Correlation Coefficient in Medical Research", *Malawi Medical Journal*, vol. 24, no. 3, pp. 69-71, 2012.
[21] T. Chai and R. R. Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?-Arguments against Avoiding RMSE in the Literature", *Geoscientific Model Development*, vol. 7, pp. 1247-1250, 2014.
[22] H. Kang, "The prevention and handling of the missing data", *Korean Journal of Anesthesiol*, vol. 64, no. 5, 402-406, 2013.
[23] R. Jabrah, *et al.*, "Using ranked auxiliary covariate as a more efficient sampling design for ANCOVA model: analysis of a psychological intervention to buttress resilience", *Communications for Statistical Applications and Methods*, vol. 24, pp. 241-254, 2017.
[24] S. A. Culpepper and H. Aguinis, "Using Analysis of Covariance (ANCOVA) With Fallible Covariates", *Psychological Methods*, vol. 16, no. 2, pp. 166-178, 2011.
[25] G. Gordon and S. Qiu, "A divide and conquer algorithm for exploiting policy function monotonicity", *Quantitative Economics*, vol. 9, issue. 2, pp. 521-540, 2018.