International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 3649~3656
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3649-3656

Estimation of regression-based model with bulk noisy data

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

Article Info

Article history:

Received Sep 18, 2018
Revised Apr 8, 2019
Accepted Apr 17, 2019

Keywords:

Bulk noise
Classification
Estimation
Mean square error
Noisy and missing data
Regression-based model

ABSTRACT

The bulk noise has been provoking a contributed data due to a communication network with a tremendously low signal to noise ratio. An appreciated method for revising massive noise of individuals through information theory is widely discussed. One of the practical applications of this approach for bulk noise estimation is analyzed using intelligent automation and machine learning tools, dealing the case of bulk noise existence or nonexistence. A regression-based model is employed for the investigation and experiment. Estimation for the practical case with bulk noisy datasets is proposed. The proposed method applies slice-and-dice technique to partition a body of datasets down into slighter portions so that it can be carried out. The average error, correlation, absolute error and mean square error are computed to validate the estimation. Results from massive online analysis will be verified with data collected in the following period. In many cases, the prediction with bulk noisy data through MOA simulation reveals Random Imputation minimizes the average error.

Corresponding Author:

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

1. INTRODUCTION

Currently, noisy and missing datasets are on the rise due to malfunction in sensor technology. These noisy parts disturb particular fields in the table dataset. Human being can figure out what these blanks should stand for, or guess around them. But most programmers think these the hard problems exactly because they have no clue about the method to apply. The threat of fixing noisy data is not only a mental fear but also turns a simple computation to an extra calculation, just only to find out where the missing data is located. Keeping this observation in mind, some approaches which will help calculation for data analytics work out while solving missing data problems must be provided so that data interpretation can be performed.

Basha et al. [1] propose a similarity-based cluster for classifying the string. It calculates the cosine similarity then compares with the vector of the information of the text. This technique is applied for text-based datasets, and the performance is investigated using signal to noise ratio, mean square error, and execution time. Lee et al. [2] propose a fuzzy-based approach to classify the sentence text in many document datasets. A fuzzy metric has been analyzed to translate the high level into a low-level text. The proposed cluster divides the state space into sub-spaces. The discrete sub-spaces are thus united to create the separate level. The approach produces optimization and runs faster than other techniques. Wei et al. [3] introduce a lexical-based sentence. It employs the ontology ordered patterns to compute the similarity between individual words. The lexical-based method is used to analyze the semantic of the words. Compared to the text clustering, the clustering of the empty (blank) texts are the most difficult. As the missing words in the dataset are widely encountered by applications in reality, the clustering of the noisy dataset needs the attention. In this paper, the proposed method based on estimation centers the case of missing data.

In real life, noisy data [4] develops whenever blank has replaced any entries in the table as shown in Figure 1. Noise data can develop and roots an unreliable consequence ranging from the original dataset cannot be executed (error found) to nonresponsive process while running the file. Liu et al. [5] introduce a tree-based clustering algorithm for classifying texts in parallel. The parallel algorithm per se minimizes the time convolution by executing the data collection, the data partitioning, and classification. The master-slave process in parallel processing is applied. The proposed method improves time complexity and accuracy. Shi et al. [6] suggest a genetic algorithm to combine the fitness equation to the convergence function in K-Means clustering. Documents based upon topics are summarized to provide a better comprehension. Li et al. [7] propose a Fuzzy-based algorithm to increase the efficiency and accuracy of the text dataset. It proves that the classification is more effective than the regular text partitioning algorithms. Bano et al. [8] develop the training algorithm based upon the neural network model. Ranking-based clustering and filtering technique are used to eliminate the distinct data.

Noisy data represents a worthless data. Any illegible data which is collected automatically, will outcome and can be described as noise. Shabir et al. [9] accord a denoise process to recover the quality of satellite image. These noises are inclusive of not only incompatible problems such as hardware or software incompatibility but also processing hazards such as no execution, no operation, rejection or failure. A little portion of noise can simply scramble the clustering process of measured data. It is more trouble in case of image data from satellite as it connects with geostationary. If an algorithm focuses on a digital transformation, it quantifies noise into error data. If it is to be seriously harsh to noise, it has to denoise then fine-tune. However, in the case of bulk noisy data, it worsens the originality and causes a great hazard.

Store	Total Price					
	Food	Drink	Vegetable	Electronics	Appliances	Stationery
A	10		4	865		12
B						27
C		28				
D						
E	58					

Figure 1. Example of noise data (blank)

The objective of this paper is to estimate the accuracy of the regression-based model for bulk noise data using MOA [10] simulation. In many cases of bulk noise data, noisy part is such a large fraction (above sixty percent) of the total data size. The reason why it's called "bulk noise" is that irrational fluctuation is commonly due to variables which are unable to be accounted for. Bulk noise will be seriously taken into account from many practical cases. Firstly, the noisy deterministic part will be inquired to break through the error in manipulation. Secondly, the proposed estimation will treat these blank data then regression-based results from simulation are used to validate the proposed method. Lastly, the precision of bulk noise treatment will be discussed after comparing with the actual dataset.

2. RELATED WORK

The noisy data is exactly fuzzy as it comprises of any value (including blank) which is unreliable and unused. It could be equipment errors, human errors, and improper data entry. Noisy data is commonly found as medical datasets are held. These values develop unknown challenges to the data scientist. A method for treating the missing values is proposed [11] as the characteristic values are naturally recurring using non-parametric discretization approach. The concept of z-score to treat these missing parts in the datasets is applied. The requirement is only that data needs to be recurred. The authors [12] propose technique of handling noise data occurred in medical datasets by classifying, estimating, and clustering. One of the vital techniques in statistics is to find the location of all deterministic parts which will be used for estimation. In many cases noisy data can be simply identified if a small fraction of the total part is missing. Refer to "R-Square" with a low portion of noisy data, the regression technique helps divide randomly and visualizes for how majority of the deterministic part can play role in the final outcome. Note that regression-based imputation is also presented in [13, 14] to figure out missing values and the accuracy is measured by utilizing MSE metric, however, it is regardless of estimation model. Noisy data can also occur widely as bulk in the dataset. In this case, the minority part is fresh value and is useable, while the majority one has to be discarded or retreated.

However, the investigation in this paper is not like other methodologies as mentioned above, firstly, the proposed treatment is employed until the bulk noise is uninvolved. After the elimination of the unwanted bulk noisy data, the proposed algorithm keeps fixing all substantial blanks in the dataset by estimating the substitution. Note that the single element of noise in the dataset can hamper the data manipulation unless exclusion of this major noisy data. Four proposed treatments are introduced, namely, Duplication (DU), Mean Variables (MV), Deletion Mechanism (DM), and Random Imputation (RI). These four algorithms are applied for replacing noise with substitution. Secondly, the computation costs are inclusive of searching time for the elimination of bulk noisy data and individual algorithm execution time. As a down-to-earth, thirdly, these four estimations are validated using the comparison with actual values in order to reflect their accuracy. Lastly, the experimental results using MOA simulation are summarized. The awaiting sections are categorized as follows: Section III explains the existing bulk noise situations. Section IV describes the performance results of these proposed techniques. Section V outlines the conclusion of the paper.

3. DATASETS WITH BULK NOISE

In this section, features of bulk noise are illustrated, datasets with majority of noise are listed and the inherent structure of bulk noise is sorted out.

3.1. Noise prevenience

A single entry of noise can corrupt a whole dataset. Noise can be easily originated from faults during data collection procedure or storing process. An elementary existence of noise stops the extraction of insightful data in data curation, which can reflect the defected machine learning operation. In practice, it can be desperately complicated to handle faults as such. Therefore, to pinpoint and to treat noise in any dataset may conquer the manipulation constraint and the impediment learning process. In this paper, thus, the severe case of bulk noise in the dataset is investigated. Bulk noise turns the presence of noise in the dataset to be beyond 50% of the whole. The complexity is to allocate and search thoroughly where the bulk noise presents. This pinpoint can provide the conclusion regarding the necessary of the bid of noise treatments.

To purge bulk noise, most techniques assume a) the dataset is exactly at hand for execution and b) the dataset is small for manipulating at a time. This is due to the implementation of a rule-based approach. But in this paper, a partition-based strategy (PBS) is adopted by giving that a dataset S can be divided into two portions: a minor part (subset) and a major (bulk) part, in which the bulky part is assumed to prompt the identification of the major noise from the whole dataset. Consequently, the assumption is realistic, especially in noisy environments. In case of huge datasets, bulk noise elimination is scaling up partition filtering time accordingly. The experiment on synthetic datasets with noise level (even above 50%) shows the effective performance after PBS.

A modest technique [15] to involve bulky values of noisy data is to remove the whole instance which specifically contains the noise data of any attributes. However, this approach does not crack the problem in case of bulk noise as, only a few or minority of these instances remains. It is feasible that these discarded instances may be critical and sensitive factor as executing data curation. To identify noise in the dataset is described. Two cases are now considered, the case where the noise is random and the case in which a bound on the noise exists. In the first one, in case of huge datasets, a synthesizer will compute feasible representations for a dataset then estimates the dataset based upon the calculation. In the latter one, optimization is expected on the simulation.

3.2. Simulation datasets

Filtering then detaching instances which are majority in bulk noise dataset is a main purpose of data curation pre-processing (both PBS filtering and anomaly detection) because these bulk noises obstruct then stop data analytics. Proposed four treatment approaches for estimating data for majority of noisy values which are Duplication (DU), Mean Variables (MV), Deletion Mechanism (DM), and Random Imputation (RI) have been introduced. Let T be a given noisy dataset matrix which contains a rows and b columns, while k represents instances affected by bulk noise or noise level, in which k is always less than a ($k < a$ and T_{k1}, T_{k2}, T_{k3},…, $T_{k(b-1)}$, T_{kb}) for each k = 1, 2, 3,…, a. The T matrix is expected to be a deterministic set. An element T_{kb} is set to be a noisy element whenever $\{T_{ij} = \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$. Note that in case of bulky noise, k is always identical or larger than the half of a ($k \geq a/2$). The dataset with bulk noise is called hampered dataset. Thus treatment techniques to overthrow the hazard and work out with the analysis by applying the estimated vector E_n are described in the next section.

The partition-based filtering techniques emphasize on discarding noise which can be detected at low-level data faults established by an impaired collection of data, but instances which are defected can bar the analytics. Noise can direct to misinterpret (negative concern), leading data analyzer to take on a relation

of any attributes (fault decision) although it legitimately is not (a type I error). Therefore, if it is to certain data analytics, these T_{kb} instances must be denoised, regardless of any analysis. It is also critical for denoise approaches to detach any noise data. In case of the bulk portion of noise where $k \geq a/2$, any methods have to neglect a remaining fraction of the whole dataset as a removal can result a certain few or only minor portion. This paper focuses on four types of the proposed treatment for bulk noise, which include Duplication (DU), Mean Variables (MV), Deletion Mechanism (DM), and Random Imputation (RI) in order to grant data analytics in reality. The experiments are based on the regressive model with ten different synthetic datasets. In the individual experiment, the estimation is computed after denoising in terms of the effect on the data analysis. In the consequent year, once the actual data is collected then, they will be compared to those early year estimations.

4. RESULTS AND ANALYSIS

The open-source based MOA is selected for the data analytics. Ten datasets are taken up and the assessment of a regression model for bulk noisy with the dissimilar noise level (k) is measured. The experiment is manipulated on an Intel® Core ™ i5 CPU, 1.60 GHz Processor and 8 GB RAM on board. The datasets are chosen in order that they all are dissimilar in size, number of instances, and attributes.

4.1. Mean square error

Mean squared error (MSE) is a metric to quantify the differences between sample and population values anticipated by a regression line or forecasted values of the observations. The lower the MSE, the closer to the line of best fit is found. The MSE explains the standard deviation of the dissimilarity between observation and prediction. The dissimilar value is calculated by the targeted data execution over the errors in estimation. MSE basis is a balance between variance and bias is shown in (1).

$$\mathrm{MSE} = \frac{\sum_{t=1}^{D}(\hat{x}_t - x_t)^2}{D} \quad (1)$$

Where, x_t is time series of the finite observation, $\hat{x}_t$ is the forecasted time series, and D denotes the number of sample data. A dataset in a training condition declines the error rate for experiment set. Fault rate for training dataset will be comparatively higher than that of the experiment set. If any two techniques provide the equal mean absolute error then MSE is taken up for deciding, which is the optimum approach.

4.2. Mean absolute error

The mean absolute error (MAE) is a figure used to quantify predictions of the critical results. The MAE is a mean of the absolute value of faults and can be defined by

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|x_k - \hat{x}_k| \quad (2)$$

Where, x_k is the finite observation time series and $\hat{x}_k$ is the predicted time series.

4.3. Noise structures

Fully at Random (FAR) defines bulk noise structures are not relating to any factors. For example, most questions ask the respondent for a random answer. Intentionally (ITT) outlines bulk noise structures added up with privacies. For instance, some respondents clumsily echo their sensitive data such as age, income, age, etc. They end up filling with a blank deliberately or white-false figures. Many bulk noise data in this paper include both ITT and FAR.

4.4. Proposed treatments

Four proposed treatments for assigning data to involve the problem of bulk noisy values which are based on duplication, mean value, deletion, and random value are described. Let T be a given noisy dataset matrix which contains a rows and b columns, where k is noise level as well as smaller than a ($k < a$ and T_{k1}, T_{k2}, T_{k3}, ⋯, $T_{k(b-1)}$, T_{kb}) for each $k = 1, 2, 3, \cdots, a$. The T matrix is expected to be a deterministic set.

An element T_{kb} is set to be a noisy element whenever $\{T_{ij} = \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$. In the case of bulky noise, k is equal or greater than the half of a ($k \geq a/2$). The T dataset is called hampered dataset.

Thus, treatment in order to get over the stoppage and move on with the analytics using the estimated vector E_n are presented as follows.

4.4.1. Duplication (DU)

Duplication involves with the dataset *T* by firstly deleting all *k* instances. This approach benefits the simplest treatment whether or not there are impacts on the elimination. After the entire *k* rows of the matrix *T* are removed, then estimated E_n dataset is $\{d_{ij} \neq \phi \parallel \infty, 1 \leq i \leq (a\text{-}k); 1 \leq j \leq b\}$. Although it seems the DU treatment nurtures an unfair estimation by using the remaining *a-k* instances to duplicate until *En* dataset grows to $\{d_{ij} \neq \phi \parallel \infty, 1 \leq i \leq a; 1 \leq j \leq b\}$ but the paper investigates a bulk noise sample and examines whether or not the DU approach reflects an acceptable strategy.

4.4.2. Mean variables (MV)

Mean value criterion is to impute data to substitute all *k* instances. Apply PBS to the targeted *T* dataset and classify a dataset which comprises of *k* instances. Any *k* rows of the matrix *T* possess an element d_{ij} with noisy data where $\{d_{ij} = \phi \parallel \infty, 1 \leq i \leq k; 1 \leq j \leq b\}$ then the entire row is replaced by employing the MV substitution for estimated E_n dataset as listed in Equation 3.

$$d_{ij} = \frac{1}{|a-k|} \sum_{x=k+1}^{a} d_{xj} \quad (3)$$

The investigation of the MV is that it is an acceptable estimation for a parameter out of a normal distribution. In case of ITT, this treatment induces a volatile bias. Not to mention the MV is led by the distorted replacement as well as develops the size of state space compared to the above DU.

4.4.3. Deletion mechanism (DM)

DM is simply removing all *k* instances. The DM claims the simple but quicker treatment whether or not the deletion will influence the future analytics. After the entire *k* rows of the matrix *T* are deleted, then estimated E_n dataset is $\{d_{ij} \neq \phi \parallel \infty, 1 \leq i \leq (a\text{-}k); 1 \leq j \leq b\}$. The DM treatment promotes humble analytics and a fair prediction in which state space is insignificant, therefore, the DM is a provoking approach.

4.4.4. Random imputation (RI)

Utilize several imputations at random for replacement. Similar to MV, the PBS is applied to the targeted *T* dataset and results a dataset with *k* instances. Any *k* rows of the matrix *T* possess an element d_{ij} with noisy data where $\{d_{ij} = \phi \parallel \infty, 1 \leq i \leq k; 1 \leq j \leq b\}$ then the entire row is substituted by using the RI replacement for estimated E_n dataset.

The minimum likelihood found in column *j* (where $j = 1, 2, 3, \ldots, b$) is marked by $d(\min)_j$ where $d(\min)_j = \text{Min}(d_{kj})$ for each $k = 1, 2, 3, \ldots, (a\text{-}k)$. Likewise, the maximum likelihood of column *j* (where $j = 1, 2, 3, \ldots, b$) is defined by $d(\max)_j$ where $d(\max)_j = \text{Max}(d_{kj})$ for each $k = 1, 2, 3, \ldots, (a\text{-}k)$. The substitution for estimated E_n dataset with multiple imputations for *k* instances in each column *j* is randomly explained as follows:

$$d_{ij} = RANDOM\left[d(min)_j, d(max)_j\right] \quad (4)$$

State space can be significant to reflect the speed of computing power. In this paper, the computation cost is summarized, according to the performance evaluation. It is apparent that any predictions are problematical if the computation cost is high as depicted in Table 1. Note that in case of bulk noise, *a* is always smaller than 2*k*.

Table 1. Computation cost of proposed treatments

Treatment	Computation Cost
DU	$O(ab) + O(ab\text{-}bk) \approx O(ab)$
MV	$O(ab) + O(ab\text{-}bk) \approx O(ab)$
DM	$O(ab)$
RI	$O(ab) + O(2(ab\text{-}bk)) \approx O(ab)$

It is proposed that treatments for involving the bulk noise can be outlined into two collective strategies, a model-based strategy (MBS), and a partition-based strategy (PBS). The PBS will divide and iron

out the bulky noise part before applying the estimation. While the MBS revises the algorithms in order to carry out the noise data before the parametric estimation is applied. The common MBS is used in PSPP or ANOVA software, which applies various imputations for replacing a noise data. PBS technique implements likelihood information into attention rather. With the MBS algorithm, the employment is complex and the skill is required because the algorithm per se has been fundamentally designed to reflect the parameter centric especially to the state spaces. If the MBS's algorithm provides common results, such as a variance or vector of average values, then it can be called the data centric. The error metrics of ten dissimilar datasets using MOA at noise level ranging from 50% to 80% will be collected. This is a simple analysis toward the designated datasets, and results are listed in Table 2-5. The three errors in the table characterize the mean squared error (MSE), the correlation coefficient (COEF), and mean absolute error (MAE) correspondingly. Dataset#5 provides lowest amount for both MAE and MSE while dataset#2, #7, #8, and #9 obtain smallest figure for COEF. The regression-based prediction is also summarized in Table 6.

Table 2. Estimation with mean squared error for ten different datasets (k = 0.5)

Dataset	MSE	COEF	MAE
1	15.7	0.47	13.1
2	1.8	0.1	1.6
3	30	0.3	25.7
4	27.8	0.15	23.5
5	0.12	0.4	0.1
6	2.55	0.6	1.8
7	22.3	0.64	19.16
8	136.1	0	114.8
9	5	0.02	3.77
10	17.68	0.02	13.4

Table 3. Estimation with mean squared error for ten different datasets (k = 0.6)

Dataset	MSE	COEF	MAE
1	16.7	0.63	13.9
2	1.59	0.27	1.34
3	29.2	0.47	25.6
4	28.2	0.16	24
5	0.12	0.42	0.1
6	2.6	0.5	1.9
7	25.3	0.64	21
8	137.55	0.15	116.1
9	5	0	4.1
10	12.8	0.5	9.5

Table 4. Estimation with mean squared error for ten different datasets (k = 0.7)

Dataset	MSE	COEF	MAE
1	19.2	0.53	15.6
2	1.78	0.2	1.48
3	33.3	0.38	29.7
4	28.8	0.28	25
5	0.12	0.4	0.1
6	2.86	0.44	2
7	32.2	0.48	27
8	127.8	0.29	102
9	5.8	0.1	4.8
10	18.41	0.14	13.14

Table 5. Estimation with mean squared error for ten different datasets (k = 0.8)

Dataset	MSE	COEF	MAE
1	24.2	0.25	20.5
2	1.75	0.11	1.55
3	44.1	0.22	38.5
4	25.3	0.57	21.1
5	0.12	0.36	0.1
6	1.7	0.76	1.24
7	46.4	0.11	32.1
8	138.5	0.32	112.4
9	5.05	0.8	4
10	4.1	0.92	1.7

Table 6. Regression-based prediction for ten datasets

		Regression-based Prediction
DATASET	1	X_9=-6.07X_4+39.85
	2	X_{11}= 0.58X_9+1.75X_5+9.4X_3+2.14
	3	X_8=-0.91X_3+8.2X_5+ 216.44
	4	X_4=0.34X_1+2.26X_5-15.18
	5	X_2=X_6+0.2X_1+0.48
	6	X_6= 0.3X_3+0.1X_7- 6.96
	7	X_5=8.63X_1+0.8X_3+ 11.97
	8	X_7= 0.09X_5+199.11
	9	X_3=3.8X_1+34.82X_5+X_4+19.7
	10	X_6= -4.7X_1-4.8X_2+X_4-1.4X_5+0.1X_7+74.2

Tables 7-10 reveal average errors for a regression-based estimation comparing to the actual data. In this paper, ten different datasets are measured and the noise level (k) is ranging from 50%, 60%, 70%, and 80% as tabularized in Tables 7-10 respectively. The results in most cases display the estimation with random imputation (RI) can minimize the average error. In addition, in case of bulk noise, the computation cost for all four treatments is approximately identical. It confirms RI is the most effective mechanism for bulk noise analysis.

Table 7. Average percentage of error for ten different datasets (k=0.5)

	k = 0.5			
Dataset	DU	MV	DM	RI
1	125.8	113.6	140.2	**105.7**
2	73.65	74.64	80.17	**72.7**
3	17.3	17.6	32.8	**16.9**
4	25.8	**25.2**	47.3	26.5
5	82.3	50.4	33.4	**23.2**
6	25.7	**22.7**	56.1	25.5
7	29.03	27.88	28.57	**25.2**
8	**24.25**	24.28	33.5	25.1
9	56.6	56.8	69	**55**
10	25.6	**22.1**	24.9	44.9

Table 8. Average percentage of error for ten different datasets (k=0.6)

	k = 0.6			
Dataset	DU	MV	DM	RI
1	128.8	**114**	146.9	115.7
2	74.4	74.2	80.2	**73.3**
3	**17.2**	17.7	36.1	18.3
4	25.6	24.9	46	**24.6**
5	83.2	44.8	33.5	**23.2**
6	24.7	**23**	63	25.9
7	28.7	27.1	27.8	**24.7**
8	24.4	**24.3**	32.1	24.5
9	56.3	56.7	67.9	**54.7**
10	24.8	**21.2**	24.2	50.3

Table 9. Average percentage of error for ten different datasets (k=0.7)

	k = 0.7			
Dataset	DU	MV	DM	RI
1	133.3	118.5	155.3	**98.8**
2	73.1	73.3	79.3	**72.3**
3	**16.8**	17.7	41	17.9
4	25.2	24.8	45.7	**22.2**
5	83.7	40.3	33.4	**23.3**
6	26.4	**24.4**	69.4	25.6
7	30.6	28.6	29.3	**25.9**
8	24.6	**24.3**	30.7	25.2
9	56.1	56.8	65.4	**54.9**
10	24.9	**21.1**	24.2	44.3

Table 10. Average percentage of error for ten different datasets (k=0.8)

	k = 0.8			
Dataset	DU	MV	DM	RI
1	151.2	126.8	189.1	**110.6**
2	69.9	69.6	79.8	**69.4**
3	**16.3**	17.3	45.3	16.7
4	25.1	24.9	47.5	**24.3**
5	84.1	34.6	35.2	**23.4**
6	27	**24.8**	79.6	27.8
7	32.7	29.3	34	**28.1**
8	24.6	**24.1**	28.8	25.1
9	55.6	56.6	62.8	**52.9**
10	25.2	**21.6**	24.8	36.3

5. CONCLUSION

In this paper, a regressive-based estimation is presented as one of the effective tools of data analytics. In practice, the analytical data after collection has stumbled upon the noisy data which reflects directly to contents found in instances or attributes in the dataset. The problem of noise values is common in countless studies and can discontinue the rest in data curation. It has been extensively taken into consideration in the case of treatments so that the ongoing analysis can be figured out. Hence, this article deeply investigates the influence of bulk noise data and the precision of individual treatment after getting over. Ten dissimilar datasets with different noise level are taken up to estimate by using regressive based models against the actual values. The average percentage of errors from four proposed treatments, namely DU, MV, DM, and RI are studied for a bulk noise filtering approach. It is proven that RI deserves the identical computation cost, can be utilized for bulk noise treatment, and outperforms than others in many cases for the estimation. Data characteristics such as covariance and kurtosis analysis of the data will be the trial of the future research.

REFERENCES

[1] M. J. Basha and K. P. Kaliyamurthie, "An Improved Similarity Matching based Clustering Framework for Short and Sentence Level Text," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 1, pp. 551-558, 2017.

[2] S. J. Lee and J. Y. Jiang, "Multilabel Text Categorization Based on Fuzzy Relevance Clustering," *IEEE Transactions on Fuzzy Systems*, vol. 22, pp. 1457-1471, 2014.

[3] V. T. Wei, Y. Lu, H. Chang, Q. Zhou and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications*, vol. 42, pp. 2264-2275, 2015.

[4] C. Enders, *Applied Missing Data Analysis*, New York: Guilford Press, 2010.

[5] G. Liu, Y. Wang, T. Zhao, and D. Li, "Research on the parallel text clustering algorithm based on the semantic tree," *Proceeding of the 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, pp. 400-403, 2011.

[6] 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.

[7] C. Li, Y. Tan, and J. Kong, "An Mahalanobis distances based text clustering algorithm," *International Conference on Automatic Control and Artificial Intelligence (ACAI)*, pp. 465-468, 2012.

[8] S. Bano, K. R. Rao and E. S. Prakash, "Partial Context Similarity of Gene/Proteins in Leukemia Using Context Rank Based Hierarchical Clustering Algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 3, pp. 483-490, 2015.
[9] M. A. Shabir and P. T. Deepali, "Satellite Image Denoising Using Discrete Cosine Transform," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 5, pp. 372-375, 2017.
[10] A. Bifet, R. Kirkby, G. Holmes and B. Pfahringer, "MOA: Massive Online Analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601-1604, 2010.
[11] G. Madhu, *et.al.*, "A Non-Parametric Discretization Based Imputation Algorithm for Continuous Attributes with Missing Data Values," *International Journal of Information Processing*, vol. 8, no. 1, 64-72, 2014.
[12] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications*, vol. 42, no.13, pp. 5621-5631, 2015.
[13] G. Madhu and G. Nagachandrika, "A new paradigm for development of data imputation approach for missing value estimation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, pp. 3222-3228, 2016.
[14] R. Thirumahal and Patil A. Deepali, "KNN and ARL Based Imputation to Estimate Missing Values," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 2, no. 3, pp. 119-124, 2014.
[15] Neeraj Raheja and V. K. Katiyar, "Noise reduction approach based on n x 1 table and XSL display method for efficient web data extraction," *International Journal of Computer Applications*, vol. 64, 2013.