International Journal of Electronic Engineering and Computer Science
Vol. 5, No. 1, 2020, pp. 1-9
http://www.aiscience.org/journal/ijeecs

An Automated Data Pre-processing Technique for Machine Learning in Critical Systems

Monica Madyembwa, Kernan Mzelikahle*, Sibonile Moyo

Computer Science Department, National University of Science and Technology, Bulawayo, Zimbabwe

Abstract

In many critical systems, the quality of data analysis is an important factor to consider particularly if the results of the data analysis contribute towards decision making. Data cleaning techniques are used during data preparation stage, before the application of data analysis techniques on a dataset. There is a strong causal relationship between quality of data preparation and quality of results in data analysis. For this reason, data cleaning techniques have a direct bearing on the quality of results from the data analysis stage. In this paper, we propose the use of intelligent data cleaning techniques as opposed to traditional deterministic methods. It is shown in this paper that the use of machine learning techniques to clean data, particularly as used for filling-in missing data, improves the quality of subsequent data analysis. Seven (7) flight-level datasets from the US Department of Transportation (Bureau of Transportation Statistics) were used to assess whether the quality of subsequent data analysis is significantly affected by the choice of a data pre-processing technique. A set of experiments were designed with an objective of conducting a comparative analysis of the performance of data analysis techniques on data prepared using different data cleaning techniques. Three (3) data analysis techniques, namely the LSTM, FFANN and RNN, were used in the comparative analysis study to determine how each of the techniques perform depending on the data cleaning technique used. The results obtained in the comparative study indicate that the use of machine learning techniques, such as BOSOM and K-means clustering, in data preparation, increases the quality of subsequent data analysis. The quality of data analysis was measured using performance metrics such as the Cross-Entropy loss and the Mean Square Error. Both assessment metrics show improved performance for each data analysis technique if data is cleaned using machine leaning methods.

Keywords

Long Short-Term Memory (LSTM), Bat Optimised Self Organised Map (BOSOM), Artificial Neural Networks, Data Pre-processing Techniques, K-Means Clustering

Received: October 31, 2019 / Accepted: January 18, 2020 / Published online: February 14, 2020

1. Introduction

Machine learning is largely concerned with the problem of learning from known past experiences using non-deterministic mathematical models such as Artificial Neural Networks. In order to achieve learning in this fashion, any learning technique requires the use of data. The challenge is that in many cases, the data is not without errors. That is, within the data, there may exist tuples that have any combination of missing data, incorrect data, out-of-format data, un-processable data characters, et cetera [1]. This normally is the source of poor processing results. Many machine learning techniques use a number of data filling methods that allow the use of means, or medians to fill in the missing data points [2]. The challenge with this approach, while it is widely accepted, is that it tends to assume that the data is normalised [2-3]. This implies that, this method distorts patterns in the data if there are any. In other words, the method assumes that a deterministic attempt at cleaning the data is sufficient for the machine learning problem [4]. There are other methods that seek to eliminate complete

* Corresponding author
E-mail address: kernan.mzelikahle@nust.ac.zw (K. Mzelikahle)

tuples that have erroneous data. The fundamental challenge with these methods is effectively that part of the data is loss [5-6]. If a significant number of tuples have erroneous data, then a statistically significant percentage of data may be lost [6]. Further, even if the data lost is not primarily significant, the subsequent analysis on the cleaned data is inherently inconclusive because some trends would have been lost during the cleaning process. For critical systems that require machine learning, tuple elimination is an unacceptable technique as a foundation for data cleaning [7]. Tuple elimination for unclean tuples may be viewed in a sense as "erroneous data avoidance". While some erroneous data may logically be determined to be so, in many cases, data that is deemed to be erroneous may simply be data that lies outside expectation, yet it represents correct instances in the field [8, 9]. By the application of deterministic data cleaning techniques, such outliers may be lost in favour of averages. Deterministic techniques are referred to as part of logical data cleaning processes, where a researcher uses the rules of the field to identify and eliminate data entries that are outside the scope [9]. While such techniques are useful in many cases, the challenge is that they do not account for the source of the data entries. Machine learning in critical systems can not afford to make such rudimentary assumptions about the nature of data [10-11]. When a data element is being eliminated from a dataset before analysis, the fundamental assumption is that the data element is not supposed to be part of the data. This fundamental assumption can be dangerous because the out-of-range data element may be representing some unexpected behaviour in the system under study. Therefore, by eliminating such a data element, the analyst is effectively missing the opportunity to observe the unexpected behaviour from the system under study.

In this paper, a technique for data pre-processing is proposed for use in analysing data from critical systems. The assumption in this study is that data measured from a critical system represents known and unknown states of the system. Of note is that data cleaning processes that modify data may conceal certain trends within the data in favour of presenting the data in a more understandable way. Further, the use of data filling techniques introduces new data elements as if the system under study produced such data [12]. Effectively, if the system under study has erratic periods where it fails to produce data, then data filling techniques cover up this behaviour. Notwithstanding the fact that some calculations are not possible if there are gaps in the data, the technique we propose in this paper seeks to introduce machine learning for data pre-processing in order to account for sub-normal data. Notwithstanding the above stated problems, the proposed technique is a data-filling technique that uses a Bat Optimised Self Organising Map to calculate data elements that are used to fill-in either missing elements in the data or correcting out-of-range data elements. Experiments to establish the effectiveness of the techniques were conducted using the Air-Online Performance data from the US Department of Transportation. In this technique, the data modification procedure is performed 30 times and data elements are averaged using the normal distribution model to account for noise.

2. Data Pre-processing Techniques

A common technique used in pre-processing data for critical systems is the binning method. In this technique, data is categorised into classes that are known to exist for the system under study [13]. The advantage of using this technique is that it reduces the effect of noise in the data. In binning the most common approach is to discretise categories using frequency [13]. By default, for each considered variable, a two-dimensional vector group is generated, with values of the category on one dimension and frequency on the other. This technique is useful in establishing stochastic patterns from the data; however, it is insufficient in extracting inherent values and patterns particularly for critical systems [14-15]. By their very nature, critical systems require significant stability, therefore, deviations from normal behaviour that may be categorised as noise under techniques such as binning may actually provide critical clues into possible instabilities within the system.

Using the Air-Online Performance data from the US Department of Transportation, it was noted that the Departure Delay attribute is one of the significant factors that can be linked to flight risk. While white noise may be associated with acceptable behaviour in flight data, it must be noted that if one entity (such as an air-line) is to be tracked throughout the data, a significant deviation may be observed, from the standard deviation, yet on the general case such deviations may not be observable. For this reason, techniques such as binning tend to burry unique trends in favour of generalised trends. In comparison to an artificially intelligent data cleaning technique (such as one proposed in this paper), it would be observed that individual trends are not lost, and they may easily be identified.

Another data pre-processing technique, commonly used for critical systems, is cluster-based data filling [16]. In this technique, data from a number of related properties is grouped into related groupings called clusters, and normalised calculations are carried out. Using these clusters, the average de-normalised values are calculated, and are used to fill in the missing data elements [17]. This technique is close to our proposed cleaning technique in this paper. However, the difference lies in that our technique uses auto-

regression to calculate suitable data-filling points for every cluster identified. Further, our technique uses an unsupervised learning technique (BOSOM) to determine clusters for which to calculate the missing data points. White noise is considered by the BOSOM technique in the determination of clusters. This improves the consideration of both outliers and generalisability. The inherent limitation with a traditional clustering-based data filling technique is that outliers tend to be disregarded [18]. In other words, the data filling elements are calculated without the consideration of outliers, thus, they are not a true representation of the data under consideration. To deal with this limitation, the proposed technique uses a Bat Optimised Self Organising Map (BOSOM); where all data elements, including outliers, are fed into the neural network for that one particular variable and clusters are calculated.

Another data pre-processing technique for data filling is regression analysis. In this technique, the idea is to develop an *N* dimensional function that may be used to estimate missing data elements [19]. If one variable is being considered, invariably, regression analysis tends to suggest a linear trend as a solution [18-19]. Fitting the missing data element as a linear regression model is not always the best solution, because it may be misleading. In fact, when using a linear regression model for data pre-processing, a pre-judgement of the nature of data would have already been done before effective analysis. Therefore, regression analysis is not the best data pre-processing model, particularly for critical systems. In fact, using regression analysis per variable (a column) has been shown in literature [20] to be perform poorly compared to even simple filling-by-means procedure. The data filling-by-means technique is a straightforward technique where a mean of a variable is calculated in the absence of the missing data elements. The mean is then used to fill in the missing points. Simple as it is, it was shown to be more effective in determining trends compared to using regression [19-20]. This was attributed to the fact that regression attempts to determine a linear function that best suits the data, and in the event that the function is complex, then regression misses the mark significantly. In contrast, the filling-by-means approach is more succinct in that it does not bother to determine the underlying function of the data [20].

Despite the limitations observed in regression analysis, particularly for data filling, there significant advantages obtainable in regard to multi-variate data pre-processing. That is; if there are multiple variables that are related, such that one depends on the other and with missing data elements, regression analysis tends to perform better [21]. In these circumstances, regression analysis provides significant advantages than the simple filling-by-means technique because it factors in dependencies.

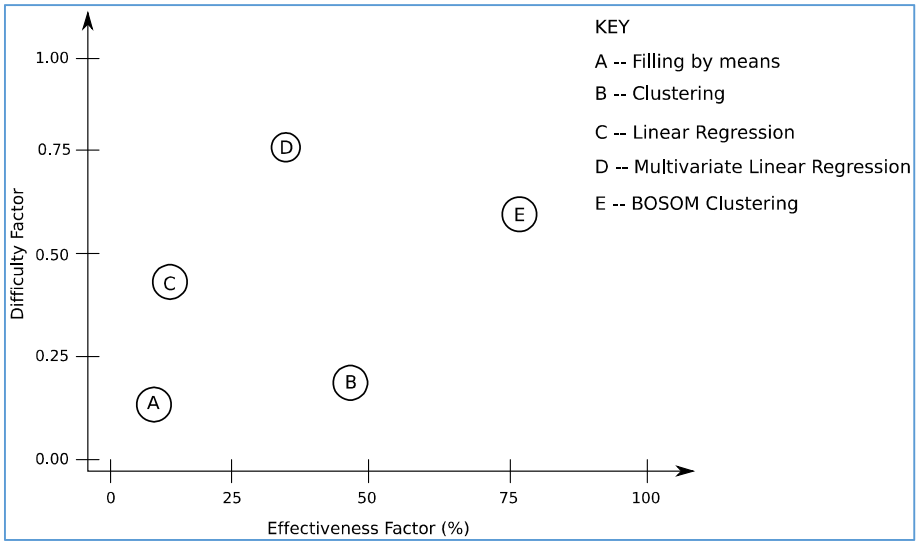

Figure 1. Effectiveness of Data Pre-processing Techniques.

In this study, an experiment was carried-out to assess the effectiveness of five (5) pre-processing techniques in respect to data analysis. The hypothesis of the study is that: The quality of data pre-processing significantly affects the quality of results in the subsequent analysis. Therefore; in this study, the results of data analyses, using a technique (for example, a Long Short-Term Memory) are contrasted against each other based on the data cleaning method used for data pre-processing. By holding the data analysis technique constant, the effect of data pre-processing can easily be interpreted. Figure 1 shows a comparison of the ease of use of the pre-processing techniques by comparing them on an applicability scale. The applicability scale evaluates difficulty of implementation against effectiveness of the technique. The observation in Figure 1 correlates to the results as shown in the Results section. As can be observed from Figure 1, the

filling-by-means technique is the simplest technique usable in data pre-processing. However, its effectiveness is limited. The clustering method is similarly simple to implement, however, it has very good effectiveness. As is shown in the Results section, the BOSOM clustering technique is highly effective, however, the implementation is somewhat difficult because there are a number of factors that require consideration. Such factors include the initialisation of the BOSOM weights and the setting up of the configuration.

3. Machine Learning Using Unclean Datasets

In order to assess the utility of the technique proposed in this paper, seven (7) datasets obtained from the air-transportation field, with a significant number of unclean tuples, were used. The datasets were prepared following a strict structured procedure in order to control for incidental results. The data is flight-level data from the US Department of Transportation, Bureau of Transportation Statistics. It is a set of logs on the arrival and departure performance of United States domestic flights operated by air carriers that are required to (or voluntarily) report to the US Department of Transportation. Different carriers may record their data differently, implying that an ontological approach is required to merge the data for similar variables. The data therefore was manually sourced and standardised. Each of the seven (7) datasets is a single file for each month of interest that includes various carries' logs. The period covered by the datasets runs from January to July of 2018.

The data elements in each dataset are variables that give the value in monetary terms, weight of shipments, number of passengers on-board, et cetera, by the origin or destination state of U.S passenger flights, exports and imports on commodity flights. Part of the data elements include sectioned data such as; flight delays, mishandled baggage, passengers on wheelchairs and, passengers with scooters. Further, there are sections that log in complaints about over-sales and airline reports on the loss, injury, and/or death of animals during air transportation.

4. The Proposed Data Cleaning Technique

In this paper we propose two (2) steps to be taken in pre-processing data for subsequent data analysis. The objective of these two (2) steps is to improve accuracy of subsequent analysis techniques. The working theory of the steps is that a data analysis technique tends to have better results if it is presented with clean data. However, many data cleaning techniques tend to distort the inherent patterns found in the data. Therefore, to eliminate the distortions, intelligent data cleaning is required. The steps proposed in this paper are summarised as follows: -

1. Use an unsupervised learning technique to cluster data into its inherent clusters for each property (e.g. for each column in the dataset). In this study, the unsupervised learning technique of choice is the Bat Optimised Self Organised Map (BOSOM).
2. Use auto-regression to fill-in missing data elements for each property based on each cluster obtained. Each property (variable or column) may be treated independently, however, for dependent variables the independent variable may be used to reflect the relationship. This approach improves the determination of fill-in values for non-linear data patterns.

To achieve these two steps, first data is manually prepared by standardising both the character-set used and the format. In formatting elements such as spacing, capitalisation, numeric data, delimitators, et cetera, a standard must be chosen and adhered to. For large datasets, automation of this process may be done by using scripts from scripting languages such as Python and AWK. In this process, care must be taken not to modify or alter the substance of the data, thus, at this stage only the formatting must be attended to. Being strict at this stage enforces the rule that pre-processing of data must not turn into data manipulation.

The Bat Optimised Self Organising Map (BOSOM) was introduced by Mzelikahle et al [22-24] for improved unsupervised learning using a Kohonen neural network. It was shown to have higher clustering abilities; thus, we chose to use it in discovering underlying patterns for each variable that has missing data points. Once clusters have been determined for each variable in the dataset, fill-in values are calculated. The auto-regression model is used to calculate the fill-in values because of its abilities to conduct non-linear forecasting. In the design of the technique, the assumption adopted is that the clusters are a collection of non-linear data with stationary data properties. In the autoregressive model used in this technique, the missing data elements were modelled as depending non-linearly on some underlying function, and on a stochastic term (an imperfectly predictable term). This way, the missing term is modelled in the form of a stochastic difference equation in relation to its closest neighbours. Suppose that a data point x_q is missing, and its neighbouring data points are x_{q-i} where q is the position of the data element on the cluster. In this case, x_q may be calculated as;

$$x_q = \frac{1}{p} \sum_{i=1}^{p} \varphi_i x_i + \varepsilon_q \quad (1)$$

where $\varphi_1, \ldots, \varphi_p$ are cluster weighted parameters, and ε_q is the white noise associated with the missing data point x_q. Using equation (1), a multiple auto-regression model is defined. Effectively, equation (1) allows forecasting of the missing data point using a non-linear combination of predictors in a cluster. In this regard, the forecast of the missing data point is dependent on a non-linear combination of neighbourhood values within the cluster for the same variable. The term auto-regression, in this paper, indicates that the missing data point is being calculated based on the regression of the variable (data column) against itself. This terminology similarly applies for a dependent variable, even when the independent variable is factored in the calculation.

The possible limitation observed in this implementation of the auto-regression is the difficulty of determining the regressive window; that is, p in equation (1). This is particularly challenging when the missing data element is at the beginning of the dataset-cluster, or at the end. It is desirable that the window determined by p be evenly spread around the missing data element x_q. In cases where x_q is at the edges of the dataset-cluster, p becomes a skewed window, either to the right or left. This skew noticeably reduces the strength of the prediction of the missing data element x_q. To address the problem relating to the skewed distribution of data points, in calculating x_q, the position of the missing data point may be randomised. Randomisation of the position of the missing data point does not affect the prediction process because the data within a cluster is considered stationary, as one of the requirements of auto-regression. The objective in randomising the position of the missing data point is to have x_q evenly surrounded within the window p. The assumption made in this process is that the data is non-linear and events that created the data entries are mutually independent.

This technique may be regarded as the simplest form of a pseudo-latent trait model. Instead of calculating the correlation between observed variables in the cluster, a canonical correlation is calculated such that the correlation between the common latent traits in a given cluster of two or more observed variables is determined. In the case of missing data points belonging to a dependent variable, the independent variable(s) are factored in using the canonical correlation technique. However, to determine the missing data point, the common variance between the two sets of variables is further standardised using orthogonal, uncorrelated components, called canonical variates. Figure 2 presents a series of steps of the data cleaning procedure based on the BOSOM technique. In the figure, the clusters are presented as individual files that are manipulated using the auto-regression data-filling technique, and they are re-merged into a cleaned dataset during the final step of the process.

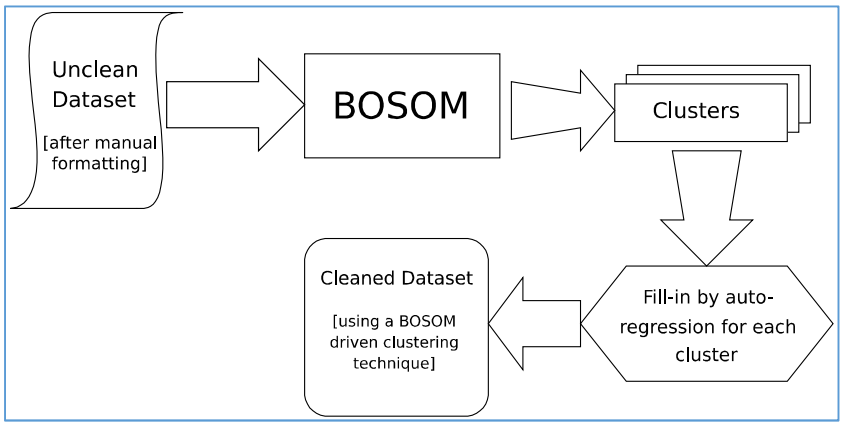

Figure 2. Data Cleaning Procedure using the BOSOM Technique.

5. Experimentation

In order to establish the utility of the proposed data cleaning technique, a number of experiments were carried out. Other cleaning techniques were used as contrasting controls such that the relative performance could be assessed. In these experiments, a total of four (4) other techniques were used to clean data and a comparative study was done. These techniques are: - (a) the filling-by-means technique, (b) the K-means clustering technique, (c) the linear regression technique and (d) the multi-variate linear regression technique. By contrasting the data cleaning methods, the relative performance of the BOSOM driven data cleaning technique was established. Three (3) analysis techniques were subsequently used to analyse the respective cleaned datasets. For each dataset, the four (4) data cleaning techniques were applied, and subsequent data analysis techniques were in-turn applied as well. The results obtained from this process were compared for each of the cleaning techniques, against results of each of the data analysis techniques. The reason for comparing multiple techniques is to eliminate incidental results that may be observed if only one technique is used. By using multiple

techniques, a generalisation of the performance of the BOSOM driven data cleaning technique was obtained. Figure 3 presents a summary of the procedure taken in setting up the experiments by following a strict set of steps. To measure the effectiveness of the data cleaning techniques, an inference is carried out based on the performance of the subsequent analysis techniques. Data unit prediction was used as the test case. That is, the cleaned data was divided into two subsets. The first subset comprised of 80% of the data, and it was used as the training set. The second set was the remaining 20% and it was used as the testing set. In this experimentation procedure, there was no data used as the validation set. Rather, the use of a comparative analysis among techniques was considered as sufficient for validation, as is the practice in literature [25]. The Mean Square Error (MSE) metric was used to calculate the performance of each subsequent technique on a given dataset, for each data cleaning method. The MSE is summarised as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \quad (2)$$

where n is the size of the population for the variable x, x_i is the data item under consideration, and $\bar{x}_i$ is the mean of the variable. Based on the test set, the MSE was used to estimate how "analysis-technique-friendly" a data cleaning method is. Another metric used for assessing the relative performance of data analysis techniques is the Cross-Entropy (CE) loss metric. In this case, the CE loss is used to measure the rate of loss of predicted data elements based on varying data cleaning techniques used. The CE metric was calculated as:

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \quad (3)$$

where, in these experiments, $H(p,q)$ is the cross entropy between variables p (the training set property) and q (the test set property). The variable x_i represents the data point on position i on a property (column in the dataset) of interest. The Cross-Entropy measures the ratio of loss during prediction; thus, the perfect CE loss is zero (0).

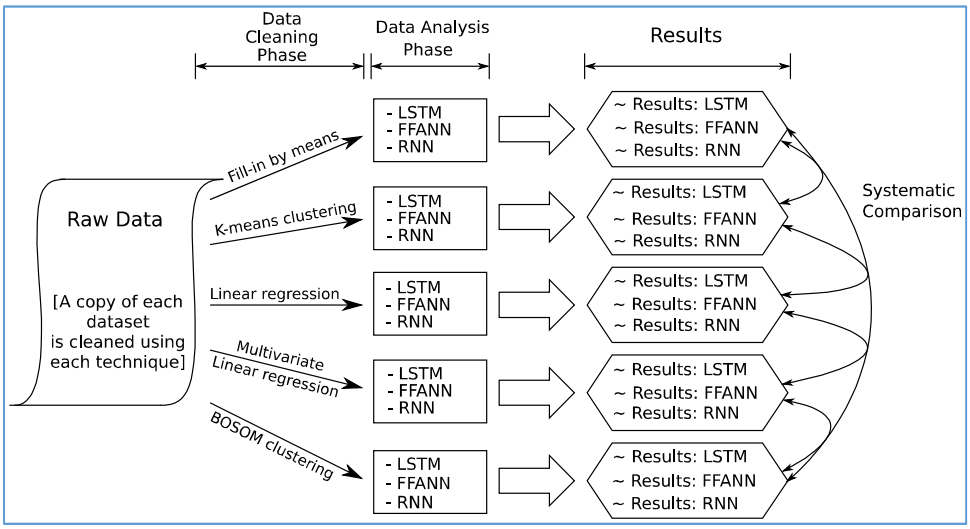

Figure 3. Summary of the Experimentation Procedure.

6. Results

The performance of the data analysis techniques (that is, the LSTM, FFANN, and RNN) was shown to improve significantly as data used for training is cleaner due to the use of intelligent cleaning methods. The problem of overfitting is not a factor within the scope of consideration on data cleaning because overfitting is related to the number of epochs used during training. The results in this paper show that for the same raw data, with the same training parameters, the performance of a data analysis technique is affected by a procedure used for cleaning the data. These results show that an intelligent data cleaning technique yields significantly better results than non-intelligent deterministic techniques. The BOSOM driven data cleaning procedure was shown to be superior in performance on measuring metrics such as the MSE, and the CE loss. These metrics both seek to reduce error on analysis techniques. When the same data analysis technique is run with data cleaned by our proposed technique, the MSE and CE loss are significantly lower compared to other cleaning techniques. These results indicate that the nature of data has a strong impact on analysis techniques in regard to their ability to detect underlying patterns. If natural noise is found in the data, and the noise does not alter the pattern of the data, then machine learning techniques such as the LSTM (Long Short Term Memory), FFANN (Feed Forward Artificial Neural Network) and RNN (Recurrent Neural Network) are not significantly impacted since they already factor for noise. However, when non-natural data elements are introduced in the data, for example during data cleaning, the risk carried is that the underlying data patterns

are altered. Notwithstanding this observation, intelligent data cleaning has the potential to preserve the underlying data patterns during the cleaning process. This observation may be seen in Figure 4, where the MSE is reduced to near zero values for BOSOM clustering-based data cleaning.

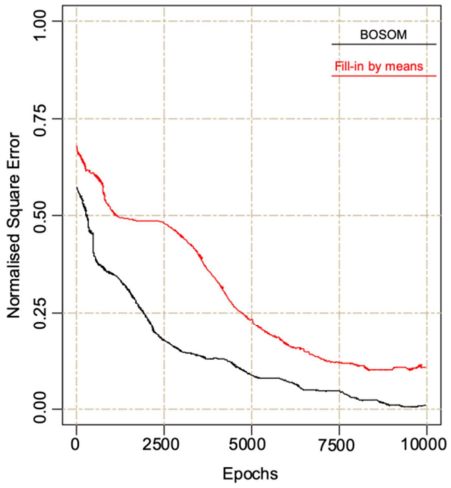

Figure 4. Comparative performance using the LSTM on data prepared by BOSOM clustering and Fill-in-by-means techniques.

In Figure 4, a comparison is made for the performance of the LSTM using data cleaned by a "Fill-in by means" technique against the same data cleaned by the BOSOM clustering method. On all the seven (7) datasets, the ability of the LSTM to determine the underlying patterns is significantly increased, hence the very low MSE values. The important observation from Figure 4 is that for all epochs, the LSTM performs better on data cleaned based on BOSOM clustering compared to the "Fill-in by means" technique. The lack of a crossing point on the two graphs imply a significant impact of using intelligent data cleaning. Figure 5 shows a comparative analysis of the performance of the FFANN on the CE metric for different data cleaning techniques.

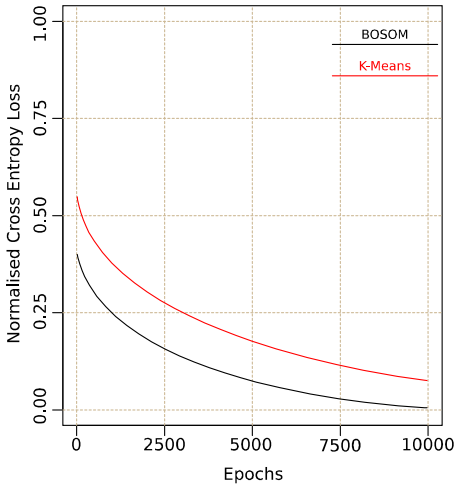

Figure 5. Comparative performance using the FFANN on data prepared by BOSOM clustering and K-means clustering techniques.

Similar to results observed in Figure 4, the results in Figure 5 reveal that the CE loss progressively decreases towards zero (0). The normalised CE loss approaches zero (0) for both data cleaning techniques when run on the FFANN, however, the rate of CE loss is better for the BOSOM based technique compared to the K-means clustering technique. Both, the BOSOM clustering technique in data cleaning and the K-means clustering technique are considered intelligent data cleaning methods. The fact that BOSOM clustering-based data cleaning significantly outperforms K-means clustering based data analysis is interesting to note, however, further experiments need to be conducted in order to draw conclusive assertions. In this study, the mere fact of significant difference in performance was sufficient to show that the quality of data cleaning has an impact on the quality of results obtainable in any data analysis technique. This argument introduces the idea that data analysis is incomplete without proper data-preparation. The second idea introduced in the paper is that intelligent data preparation may provide better results compared to traditional deterministic and numerical methods. Figure 6 shows a comparison of performance between data prepared using BOSOM clustering and Multi-variate linear regression on Recurrent Neural Networks (RNN) analysis. Similar to the analyses done using the LSTM and FFANN, the RNN analysis was used for trend analysis and forecasting across all the seven (7) datasets.

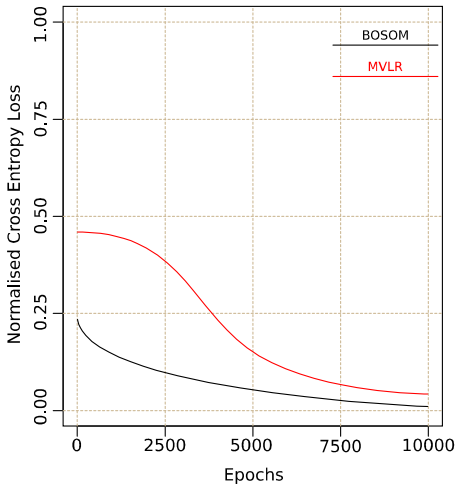

Figure 6. Comparative performance using the RNN on data prepared by BOSOM clustering and Multi-variate Linear Regression (MVLR) techniques.

Figure 6 shows that the CE loss performance of the RNN on data prepared using BOSOM is significantly better that the performance of the same technique on data prepared using Multi-variate Linear Regression (MVLR). This finding is consistent with results observed for the LSTM and the FFANN under similar conditions. It may therefore be hypothesised that the use of intelligent data pre-processing techniques improves the quality of results for subsequent data analysis methods. Further, it may be inferred that the use of

unsupervised learning techniques in data pre-processing steps improves the chances of cleaning data without distorting inherent data patterns. Traditional numerical and stochastic techniques such as the "Fill-in by means" and the Multi-variate Linear Regression technique appear to be inferior compared to machine learning methods such as BOSOM and K-means clustering in data pre-processing.

7. Conclusion

In this paper, there is substantial evidence presented to suggest that the use of artificially intelligent methods in data cleaning and data pre-processing may improve data analysis efforts. The datasets used were extracted from the air travel case study, and it was noted that more accurate determination of data patterns may be extremely important in critical systems. Predictions and forecasts are of paramount importance, particularly if used as part of safety measures in critical systems. For this reason, the data pre-processing stage needs improvement in order to enhance the quality of data analysis.

References

[1] Demuth, H. B., Beale M. H, De Jess, O. and Hagan, M. T. (2014). *Neural Network Design*. 2nd ed., Martin Hagan, Oklahoma, USA: Oklahoma State University.

[2] Graupe, D. (2007). *Principles of Artificial Neural Networks*. 2nd Ed. New Jersey, USA: World Scientific.

[3] Kröse, B. J. and van der Smagt P. (1996). *An Introduction to Neural Networks*. 8th ed. Department of Computer Systems, University of Amsterdam, Netherlands.

[4] Batista, G. E., Prati, R. C. and Monard, M. C., 2004. *A study of the behavior of several methods for balancing machine learning training data.* ACM SIGKDD explorations newsletter, 6 (1), pp. 20-29.

[5] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S. and Xin, D., 2016. *Mllib: Machine learning in apache spark*. The Journal of Machine Learning Research, 17 (1), pp. 1235-1241.

[6] Winkler, W. E., 2003, August. Data cleaning methods. In Proc *ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation.*

[7] Ng, H. W. and Winkler, S., 2014, October. A data-driven approach to cleaning large face datasets. In: *2014 IEEE International Conference on Image Processing* (ICIP) (pp. 343-347). IEEE.

[8] Maloof, M. A. ed., 2006. *Machine learning and data mining for computer security: methods and applications.* Springer Science & Business Media.

[9] Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J. and Wu, E., 2016, June. Activeclean: An interactive data cleaning framework for modern machine learning. In: *Proceedings of the 2016 International Conference on Management of Data* (pp. 2117-2120). ACM.

[10] Cazorla, L., Alcaraz, C. and Lopez, J., 2013, September. Towards automatic critical infrastructure protection through machine learning. In: *International Workshop on Critical Information Infrastructures Security* (pp. 197-203). Springer, Cham.

[11] Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A. and Clifford, G. D., 2016. *Machine learning and decision support in critical care*. Proceedings of the IEEE. Institute of Electrical and Electronics Engineers, 104 (2), p. 444.

[12] Varshney, K. R., 2016, January. *Engineering safety in machine learning.* In: 2016 Information Theory and Applications Workshop (ITA) (pp. 1-5). IEEE.

[13] Naeini, M. P., Cooper, G. and Hauskrecht, M., 2015, February. *Obtaining well calibrated probabilities using bayesian binning.* In: Twenty-Ninth AAAI Conference on Artificial Intelligence.

[14] Rahimi, A. and Recht, B., 2008. *Random features for large-scale kernel machines*. In: Advances in neural information processing systems (pp. 1177-1184).

[15] Murray, J. F., Hughes, G. F. and Kreutz-Delgado, K., 2005. *Machine learning methods for predicting failures in hard drives: A multiple-instance application.* Journal of Machine Learning Research, 6 (May), pp. 783-816.

[16] Xu, B. and Chen, D. Z., 2007, May. Density-based data clustering algorithms for lower dimensions using space-filling curves. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 997-1005). Springer, Berlin, Heidelberg.

[17] Xue, G. R., Lin, C., Yang, Q., Xi, W., Zeng, H. J., Yu, Y. and Chen, Z., 2005, August. Scalable collaborative filtering using cluster-based smoothing. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 114-121). ACM.

[18] Agyemang, M., Barker, K. and Alhajj, R., 2006. *A comprehensive survey of numeric and symbolic outlier mining techniques*. Intelligent Data Analysis, 10 (6), pp. 521-538.

[19] Vazhkudai, S. and Schopf, J. M., 2003. *Using regression techniques to predict large data transfers*. The International Journal of High-Performance Computing Applications, 17 (3), pp. 249-268.

[20] Kotsiantis, S., Kostoulas, A., Lykoudis, S., Argiriou, A. and Menagias, K., 2006, July. *Filling missing temperature values in weather data banks*. In: 2006 2nd IET International Conference on Intelligent Environments-IE 06 (Vol. 1, pp. 327-334). IET.

[21] Chen, B. W., Rho, S., Yang, L. T. and Gu, Y., 2018. *Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities*. Future Generation Computer Systems, 78, pp. 859-866.

[22] Mzelikahle, K., Mapuma, D. J., Hlatywayo, D. J. and Trimble, J., 2017. Optimisation of Self Organising Maps Using the Bat Algorithm. *American Journal of Information Science and Computer Engineering*, 3 (6), pp. 77-83.

[23] Mzelikahle, K., Trimble, J. and Hlatywayo, D. J., 2018. A Hybrid Technique Between BOSOM and LSTM for Data Analysis. *International Journal of Mathematics and Computational Science*, 4 (4), pp. 128-138.

[24] Mzelikahle, K., Hlatywayo, D. J. and Trimble, J., Application of the BOSOM-LSTM Technique in Seismic Vulnerability Assessment. *American Journal of Geophysics, Geochemistry and Geosystems*, 5 (1), pp. 29-39.

[25] Gers, F. A., Schraudolph, N. N. and Schmidhuber, J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, 3, pp. 115–143.