A NOVEL APPROACH TO OUTLIERS REMOVAL IN A NOISY NUMERIC DATASET FOR EFFICIENT MINING

A. R. Ajiboye[1], K. S. Adewole[1], R. S. Babatunde[2] & I. D. Oladipo[1]
[1]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer, Library & Information Science, Kwara State University, Malete, Nigeria.
[1]ajibabdulraheem@gmail.com ; [1]adewole.ks@gmail.com;
[2]ronkebabs711@gmail.com; [1]idoladipo@gmail.com

ABSTRACT

Data pre-processing is a key task in the data mining process. The task generally consumes the largest portion of the total data engineering effort while unveiling useful patterns from datasets. Basically, data mining is about fitting descriptive or predictive models from data. However, the presence of outlier sometimes reduces the reliability of the models created. It is, therefore, essential to have raw data properly pre-processed before exploring them for mining. In this paper, an algorithm that detects and removes outliers in a numeric dataset is proposed. In order to establish the effectiveness of the proposed algorithm, the clean data obtained through the implementation of the proposed approach is used to create a prediction model. Similarly, the clean data obtained through the use of one of the existing techniques is also used to create a prediction model. Each of the models created is simulated using a set of untrained data and the error associated with each model is measured. The resulting outputs from the two approaches reveal that, the prediction model created using the output from the proposed algorithm has an error of 0.38, while the prediction model created using the cleaned data from the clustering method gives an error of 0.61. Comparison of the errors associated with the models created using the two approaches shows that, the proposed algorithm is suitable for cleaning numeric dataset. The results of the experiment also unveils that, the proposed approach is efficient and can be used as an alternative technique to other existing cleaning methods.

KEYWORDS: *Algorithm, Data mining, Data pre-processing, Outliers, Prediction.*

1. INTRODUCTION

Outliers are observations that has high degree of deviation from the majority of the observation data (Liu et al., 2004). In the process of data capturing, there are a number of factors that may make the outliers to find its way into a dataset. Studies have shown that some of these factors are sometimes avoidable. Prominent among them according to the report in (Hodge & Austin, 2004), include faults that can be traced to mechanical, sometimes it may be as a result of system or fraudulent behaviour; others may be due to instrument or human errors. Outliers which include a certain amount of exceptional values, may also be identified in dataset as what can simply be seen as natural deviations in populations. The dataset that is free of noise produce better, promising and reliable results than the set of data that is filled with anomalies.

In order to get rid of inconsistencies in the dataset, the current approaches used in the cleaning of such data typically follow three main components: auditing of the data to identify discrepancies, choosing transformations to fix them, and applying the transformations on the dataset (Raman & Hellerstein, 2001). Outlier detection has extensive use in a wide variety of applications, the study in (Singh & Upadhyaya, 2012) listed some of these applications as: military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems.

In the process of making efforts to clean dataset, sometimes, it may be necessary to merge two or more files stored in different databases. This task may be inevitable in order to have them stored in a single location for easy exploration. Removal of outliers or cleaning of the noisy dataset to make them suitable for mining is a key step in data mining process (Han et al., 2012); and in order to detect the inconsistencies in the data, some exploratory methods such as: Statistical, clustering, pattern-based and association rules were identified in (Ahmed & Aziz, 2010). The study also identifies a framework for effective handling of noisy data. After careful study of the use of clustering method, an enhancement is introduced into

the procedures it follows through the algorithm proposed in this study. This has resulted in a more efficient means of removing outliers.

The removal of outliers or anomalies in dataset has been generally found to consume the largest portion of the total data engineering effort (Zhang et al., 2003). But the benefits of such removal are enormous. One of the advantages of removing outliers in dataset is that, it eliminates redundancy. Redundancy has to do with those data that does not make any meaningful impact and capable of slowing down the processing time due to the space they occupied. The use of clustering method partitions the dataset into clusters, such that, similar and dissimilar data are separated. However, the removal of outliers using this method still leaves the blank spaces from where the outliers are removed unfilled; the objective of this study, therefore, is to design an algorithm capable of removing outliers from unclean dataset and subsequently filling up their left-over blank spaces using a standardized value.

This paper focuses on the detection and removal of outliers from the numeric structured data set. The algorithm proposed in this study is designed to support high dimensional data irrespective of the record size. The rest of the paper is organized as follows: In the next section, the related studies are reviewed and in section 3, the material used and the proposed method in this study is illustrated and discussed. In Section 4, the findings of implementing the proposed and existing algorithms on the unclean dataset are discussed, while Section 5 concludes this paper.

2. REVIEW OF RELATED STUDIES

Several studies have been reported in the literature on the approaches used to get rid of outliers in the dataset. Sometimes, one of the techniques that is usually adopted by a person or team involves reading through a set of records in order to ascertain their correctness. Although, some spelling errors are corrected during this process, while both incomplete and missing entries are also completed. However, this approach is only suitable when dealing with very few data, as it can cause monotonous

within a very short time. At times, it may possibly lead to some new errors. But in complex situations, data cleaning can be achieved by software automatically (Ahmed & Aziz, 2010).
In order to increase the efficiency of set of the data been explored and specifically, to eliminate the likely issues that might be resulted due to redundancy in the data, a data cleaning system was deployed in (Ebaid et al., 2013). To achieve a clean data, the approach proposed a series of customization of data definition to improve the quality of the data. The algorithm was reported to have improved the efficiency of the resulting outputs of the explored clean data.

The study in (Loureiro et al., 2004) proposed an outlier detection technique that is based on hierarchical clustering methods. The choice of clustering methods was reported to have been motivated by the unbalanced distribution of outliers versus normal cases in these data sets. In the study, attempts were made to create the initial clusters, but the study observed that the non-hierarchical clustering technique spreads out the outliers across the number of clusters formed. Given that most of the partitioning methods strongly depend on the initialization of the clusters, the study considered the partitioning approach as unstable.

The hybridization of a number of techniques for the cleaning of dataset has also been reported. The cleaning of data using hybrid approach proposed in (Paul et al., 2012) was to ensure a more robust data cleaning. The integration strengthens the cleaning performance of the algorithm as the features in the individual approach are maximally explored.

Study reported in (Sim & Hartley, 2006), proposed the removal of outliers using the $L\infty$ Norm. The study show that, one of the popular methods of removing outliers which involves solving an optimization problem and then discard the measurements with maximum residual is effective. However, the study opined that, the approach sometimes do not work for general minimization problems.

Also, the use of two-phase clustering process for outlier detection was proposed in (Jiang et al., 2001). The study modified k-means algorithm prior to using the method for clustering and constructed a spanning tree. The authors concluded in their findings that, the small clusters and the tree with less number of nodes were regarded as outliers.

The study reported in (Last & Kandel, 2001) proposed an automated detection of outliers in real-world data. The study was based on modeling the human perception of exceptional values by using the fuzzy set theory. The approach involves the use of separate procedures that were developed in order to detect the outliers in discrete and continuous univariate data.

The use of attribute selection was demonstrated in (Mohamed et al., 2011), the approach reduces the dataset size in the course of removing the irrelevant or redundant attributes, otherwise referred to as dimensions. The goal of attribute subset selection is to find a minimum set of attributes such that, the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

Also, study in (Low et al., 2001) proposed a knowledge-based approach for the elimination of duplicate in dataset. In the study, a transitive closure under uncertainty for dealing with the merging of groups of inexact duplicate records is computed. Findings from the study reveal that, the approach is suitable especially for identifying the duplicates and anomalies in the data set with high recall and precision.

3. MATERIAL AND METHODS

3.1 Data Collection

In order to determine how efficient the proposed approach is, this study implements the algorithm represented in Figure 1, using a public dataset. The numeric structured dataset was downloaded from an online open repository of the World Bank. The data comprised of the annual Gross

Domestic Product (GDP) from year 2000 to 2012 as retrieved from the open repository, (Databank, 2012). The data consists of 247 records, and has several missing values. In order to properly establish the effectiveness of the proposed algorithm, a number of alphabetical data and special characters were deliberately introduced to increase the presence of varieties of the anomalies in the dataset. It is these inconsistencies in the data that the proposed algorithm is designed to detect and remove.

3.2 **The proposed algorithm**

As shown in Figure 1, the algorithm proposed in this study consists of three sections: the input, output and the technique used. At the input section, a database was created and used to store the unclean data. The technique used, as shown in the algorithm involves thorough exploration of the imported data to know which of the data is valid and which ones are invalid. What is classified as valid data here, are basically numeric values, which refers to the old value in the algorithm.

Also in this paper, after the outliers has been detected and removed, a standard deviation is computed using the values within the fields where a missing value is detected. The computed value is used to replace the vacuum already created as a result of the removal of the outliers. There are several known standard values reported in the literature, which can be used in an instance like this to replace the unknown values. Using the value that appears to be the most probable to fill the missing value appears to be the most popular view as reported in (Han et al., 2012); such probable values include: mean, standard deviation, mode and other measures of central tendency.

The standard deviation is the choice in this study because, it is a measure of how numbers are spread out and it gives the square root of the variance. The computation of the standard deviation, therefore, encompasses the mean and the variance. The computation of the standard deviation is based on the formula represented in Equation 1.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad (1)$$

where δ is the standard deviation, N is the size of the data, x_i is each data in a field and

$\bar{x}$ is the mean.

Input: Unclean numeric dataset

Output: A clean and complete dataset

Technique

Step 1: Function found_data (value, outlier)
explore through all the data in each field
IF found_data = number {
return (value);
ELSE
remove (outlier) // delete the data found except number
select the field containing found_data! = number
change the selected fields to array // for easy referencing
transpose the data

Step 2: Function replace_data (oldvalue, newvalue) // replace the missing value
let D be the data in a field
verify oldvalue for nullity // check if data is missing in a field
IF yes {
newvalue = stdev (D) //compute standard value to replace missing data

```
        return newvalue;              // replace such missing data with the
        new stdev computed
        ELSE                          // but for those fields that contain
valid data
        return oldvalue;              // leave the value in their present
form
        display the new complete dataset;    // display the clean and
complete dataset
                }
        }
```

Figure 1. The proposed data cleaning algorithm

4. IMPLEMENTATIONS

The algorithm proposed in this study is implemented using PHP language, while MySQL, served as the back-end. The file that comprised of the unclean data is the input as indicated in the algorithm, relevant PHP codes are implemented on it to give the desired outputs. When outliers are removed, a vacuum is created which still need to be filled, in order to make the dataset to be suitable for exploration. Usually, several learning algorithms such as k-mean, dbscan and others do not support data with missing values. The proposed algorithm computes the standard deviation of the data around the neighbourhood of the missing values in order to give a complete clean dataset.

Clustering is an exploratory data analysis technique that grouped different objects in such a way that the degree of association between the two objects is maximized if they belong to the same cluster and minimized if otherwise. K-medoids is one of the clustering algorithms capable of partitioning the numeric data or categorical data into clusters. The algorithm is designed to cluster both numeric and non-numeric data, thus, it is an extension of k-means paradigm. It takes care of some inherent limitation of k-means, especially its inability to handle alphanumeric data is addressed in k-medoids. The algorithm clusters categorical data by

using its simple matching dissimilarity measure for categorical objects (Kaufman & Rousseeuw, 1990).

Figure 3 shows the setup of implementing the k-medoids algorithm using the RapidMiner software. The experiment separates the similar and dissimilar data into different clusters using the clustering technique. RapidMiner is a data mining application tool with many inbuilt algorithms. Like most other clustering algorithms, k-medoids does not tolerate missing values, but if value such as zero is used to replace the missing values, then the algorithm would be implemented on the dataset successfully. Figure 2 illustrates the k-medoids algorithm.

1. Choose the initial medoids
2. Determine what should be the new medoid of each cluster to update medoids
3. Assign each object to the nearest medoid
4. Compute sum of distance from all objects to their medoids
5. Repeat step 2 until the sum remains constant.

Figure 2. The k-medoids algorithm

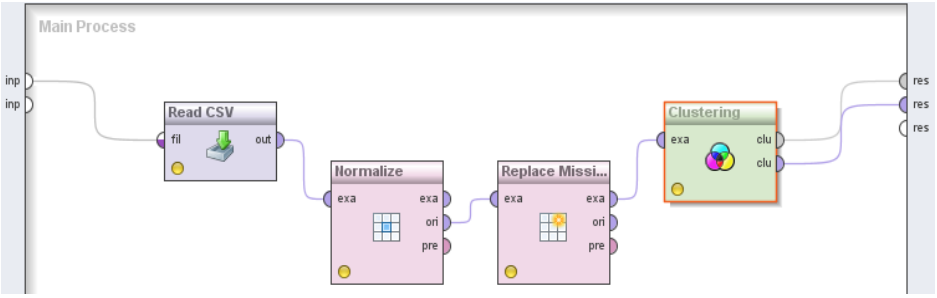


Figure 3. Implementation of k-medoids algorithm

Both the proposed approach and the implementation of the k-medoids algorithm remove the outliers from the original dataset. However, the blank spaces created as a result of the removal has left a number of blank spaces with the use of the clustering algorithm, while the proposed algorithm have them replaced with the standard value computed for the missing value in each field.

The sets of clean data produced through the two approaches were further modeled for prediction purposes. Since the data consist of GDP of several countries around the globe from 2000 to 2012, we focused on using the data for the previous five years to fit a model that is capable of predicting the average of these predictor attributes. Thus, five predictors were modeled to produce one predicted output, i.e. the average of all the five input attributes. The data are trained using back-propagation algorithm. The network predictive model created using the feed-forward neural network structure is as shown in Figure 4, and the configuration structure conforms to the parameter settings shown in Table 1. Other settings outside the listed ones are left in their default.

Table 1. The network configurations

Network Settings	
Network type	Feed-forward BP
Training function	Trainlm
Performance function	MSE
No. of Layers	2
Number of Neurons	8 (in the first hidden layer)
Transfer function	Tansig
Epochs	500
Validation check	6

Both network models have similar structure, since they have the same input, hidden and output layers, therefore, to avoid repetition of the same structure, only one of them is represented here as shown in Figure 4.

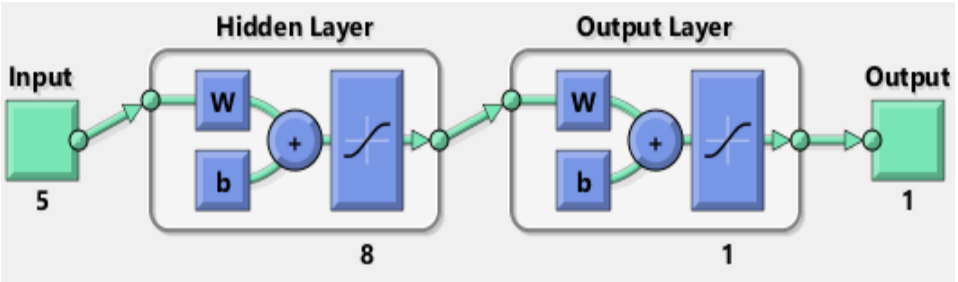


Figure 4. Predictive network model

In order to know which of the two models perform better, there is need to evaluate them. Such evaluation involves simulating each network model using untrained data. Out of the 247 records in the original data set, 67 of the records were set aside for the purpose of simulating the models created. The evaluation measures that can be used to compute the accuracy of a numeric prediction as enumerated in (Witten et al., 2011) includes: Mean Square Error (MSE), Mean Absolute Error (MAE), Relative Squared Error (RSE) and Relative Absolute Error (RAE). In this paper, the errors associated with each model are computed using the formula of Mean Absolute Error (MAE) shown in Equation 2.

$$\text{MAE} = \frac{|p_1 - k_1| + \ldots + |p_n - k_n|}{n} \qquad (2)$$

where p represents the target output, while k is the predicted output and the value of n is the size of the data. According to (Witten et al., 2011), MSE tends to exaggerate the effect of outliers in the dataset, most especially when the prediction error appears to be larger than the others. But this is usually not the case when the error is determined through MAE, as all sizes of error are treated evenly according to their magnitude.

5. DISCUSSION OF FINDINGS

The clean dataset obtained as a result of implementing the proposed algorithm and the use of clustering approach for the removal of outliers in a numeric dataset was modeled for prediction purposes. The network

models created using the techniques of neural network is evaluated using the untrained data set aside earlier for the purpose. The error associated with the two network models is as shown in Table 2. The list of the cleaned numeric data which comprised of 180 records that were used to create the network model is not shown here, due to space constraints.

Table 2. Results of simulating the two network models

Approaches	Errors
The proposed algorithm	0.38
The k-medoids	0.61

The effectiveness of each model created using the resulting outputs of the proposed and existing technique is represented in Table 2. After the two approaches have been used to create prediction models, the table shows the evaluation results which is expressed in terms of error values. The error associated with the prediction model that is created using the proposed algorithm is lower. This implies that, the model is more accurate compared to using the clustering method for the removal of outliers. The proposed algorithm also supports filling of the missing values using the computation of standard value to replace the missing data. Such replacement with valid and standard data increases the predictor data used for training.

Moreover, most learning algorithms does not support dataset with missing values, some additional tasks therefore, needs to be performed on such data to make the learning algorithm implementable on them. But the proposed algorithm is designed, such that, it can be implemented on a dataset that contain missing value and other outliers.

Also, while the clustering algorithm only partitions the original unclean dataset into a distinct cluster, the proposed algorithm detects and completely removes them. It then replaces the removed outliers with the

result of the standard deviation computed. The missing value is, thus, replaced by following this procedure for all the fields in the dataset.

The focus of most of the existing methods is on the deletion of redundant data. The proposed approach and the clustering methods implemented in this study detect and remove the outliers. In addition, the proposed algorithm is capable of replacing what is detected as invalid data. The clustering method clusters, one or two attributes at a time, it then requires a substantial period of time to complete the removal of outliers from dataset of high dimension. The proposed algorithm takes the whole dataset at once; it is therefore, faster and more accurate.

6. CONCLUSIONS

In this paper, the algorithm proposed is aimed at cleaning a noisy numeric dataset. A dataset is said to be noisy when it contains some inconsistencies, this may be in the form of missing values, or those data that are seen as natural deviations in populations. It is therefore necessary to pre-process such data before an exploratory or learning algorithm is implemented on them. This is necessary in order to reveal the patterns and other useful information embedded in the data. Outlier detection and removal is a pre-processing task in the data mining process. Other tasks that should be performed in the course of pre-processing data include: attribute selection, removal of outliers, normalization and data transformation. However, the focus of this study is the removal of outliers. Cleaning of the dataset is very crucial in the data mining process because, real world data are sometimes noisy, inconsistence and incomplete.

Although, a number of techniques that is used for the building of models from data may tolerate some level of inconsistencies in the data; however, it is definite that efforts at understanding and improving the data quality typically would yield quality patterns and better output results. The essence of data cleaning is, therefore, to ensure that those feature qualities that data must fulfill are met. Such qualities include: accuracy, integrity, completeness, validity, schema conformance, uniformity and uniqueness.

The proposed algorithm is implemented using a scripting language, PHP. The proposed approach is found to be efficient and a suitable means of removing outliers in a numeric structured data set.

REFERENCES

Ahmed, I., & Aziz, A. (2010). Dynamic approach for data scrubbing process. *International Journal on Computer Science and Engineering, 2*(02), 416-423.

Databank. (2012). World Bank Development Indicator, Annual GDP Growth. from http://databank.worldbank.org/data/reports.aspx?Code=NY.GDP.MKTP.KD.ZG&id=af3ce82b&report_name=Popular_indicators&populartype=series&ispopular=y; accessed on December 15, 2015.

Ebaid, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., Quiane-Ruiz, J.-A., Tang, N., & Yin, S. (2013). NADEEF: A generalized data cleaning system. *Proceedings of the VLDB Endowment, 6*(12), 1218-1221.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining, southeast asia edition: Concepts and techniques*: Morgan kaufmann.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.

Jiang, M.-F., Tseng, S.-S., & Su, C.-M. (2001). Two-phase clustering process for outliers detection. *Pattern recognition letters, 22*(6), 691-700.

Kaufman, L. R., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*: John Wiley & Sons Inc.

Last, M., & Kandel, A. (2001). *Automated detection of outliers in real-world data.* Paper presented at the Proceedings of the second international conference on intelligent technologies.

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers and Chemical Engineering*.

Loureiro, A., Torgo, L., & Soares, C. (2004). *Outlier detection using clustering methods: a data cleaning application.* Paper presented at the Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector.

Low, W. L., Lee, M. L., & Ling, T. W. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems, 26*(8), 585-606.

Mohamed, H. H., Kheng, T. L., Collin, C., & Lee, O. S. (2011). *E-Clean: A Data Cleaning Framework for Patient Data.* Paper presented at the Informatics and Computational Intelligence , ICI Conference, 2011.

Paul, A., Ganesan, V., Challa, J. S., & Sharma, Y. (2012). *HADCLEAN: A hybrid approach to data cleaning in data warehouses.* Paper presented at the Information Retrieval & Knowledge Management (CAMP), 2012.

Raman, V., & Hellerstein, J. M. (2001). *Potter's wheel: An interactive data cleaning system.* Paper presented at the VLDB.

Sim, K., & Hartley, R. (2006). *Removing outliers using the Linfty norm.* Paper presented at the Computer Vision and Pattern Recognition, IEEE Conference, 2006.

Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues, 9*(1), 307-323.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *DATA MINING Practical Machine Learning Tools and Techniques* (3rd Edition ed.): Morgan Kaufmann.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence, 17*(5-6), 375-381.