Vol 8. No. 1 Issue 2 – May, 2015

African Journal of Computing & ICT

www.ajocict.net

An Improved Technique for the Removal and Replacement of the Inconsistencies in Numeric Dataset

J. Abdul-Hadi
Department of Mathematics,
Bauchi State University, Gadau , Nigeria.
jamcy98@gmail.com

A.R. Ajiboye
Department of Computer Science,
University of Ilorin, Ilorin, Nigeria.
ajibabdulraheem@gmail.com

A. Abba,
Department of Statistics,
Abubakar Tafawa Balewa University, Bauchi, Nigeria.
abdulhafeezabba@gmail.com

ABSTRACT

The task of ensuring the removal of anomalies in an unclean numeric dataset, with a view to putting the data in a suitable format for exploration purposes is a major phase in the data mining process. In the process of exploring an unclean numeric dataset to unveil their useful patterns or structure, a thorough pre-processing task is inevitable in order to achieve a noise-free dataset. Poor quality data can be misleading if analysed or used to build models, hence, there is need to remove discrepancies that may be present in the data prior to exploring them. In this paper, a cleaning algorithm is proposed and implemented in order to remove the inconsistencies in a numeric dataset. The implementation of the proposed algorithm uses the Java language and the resulting outputs reveal the efficiency of the proposed approach. In order to evaluate the effectiveness of the proposed algorithm, it is compared to one of the existing methods based on some metrics. The comparisons show that, the proposed technique is efficient and can be used as an alternative technique for the removal of outliers in numeric data. This approach is also found to be reliable as it consistently gives an accurate output that is free of outliers.

Keywords: Data cleansing, Data mining, Outlier detection, Clustering.

African Journal of Computing & ICT Reference Format:
J. Abdul-Hadi., A.R. Ajiboye & A. Abba (2015): An Improved Technique for the Removal and Replacement of the Inconsistencies in Numeric Dataset. Afr J. of Comp & ICTs. Vol 8, No. 1, Issue 1. Pp 39-44

1. INTRODUCTION

Pre-processing is the task performed on the dataset in order to make it suitable for exploration. Data cleansing, data cleaning and data scrubbing are sometimes used interchangeably to describe the pre-processing task of putting the data in a clean state [1]. The real world data are sometimes incomplete or noisy and it is very rare to get a perfect data. Exploration or analysis of unclean dataset has every tendency to give a result that deviate slightly from what supposed to be the actual results. This is because the presence of anomalies in the data is capable of influencing the resulting outputs. As reported in [2], the use of quality data is crucial to getting high-quality patterns.

Putting several files together can ease exploration processes, as efforts to reveal the patterns and structure of the data would be more focused on a single database. However, integration of files from different sources is prone to duplication of records, and human errors in the course of entering data may sometimes violate the declared integrity constraints [3]. Some of the basic tasks that is performed in the process of preparing data generally involves correcting any errors typically emanates from human and/or machine input, filling in nulls and incomplete data. Manually filling of the missing value would, however, cause monotonous within a very short time, which may also lead to some new errors.

A number of methods have been reported in the literature for the detection and removal of anomalies in the dataset. Some exploratory methods identified in [4], includes: statistical, clustering, pattern-based and association rules. One of the clustering algorithms, dbscan, is implemented on the dataset explored in this study and its effectiveness is compared to the proposed approach. In a numeric dataset, zero is not regarded as missing value because it may be the expected answer to a question. The proposed approach shows a popular view of how the missing value can be replaced. The proposed approach is found to be efficient in the removal of outliers and other inconsistences in a numeric dataset.The proposed algorithm is des

igned for optimal performance at ensuring clean set of data from a set of inconsistence and noisy dataset. Unlike cleaning of few dataset, cleaning thousands of records with several unclean data require an algorithm that is capable of searching the unclean data thoroughly in order to spot all the anomalies. This is achieved through the algorithm proposed in this paper.

The rest of the paper is structured as follows: In the next section, some related studies are reviewed. Section 3 shows the proposed algorithm and discusses how it is implemented. In Section 4, the resulting outputs of this study are discussed and comparisons of the proposed approach with other existing methods based on a number of metrics are shown. The paper is concluded in Section 5.

2. RELATED WORK

Some research efforts made at improving the quality of the data as reported in the literature is discussed in this section. The study in [5], focuses on the required steps of data pre-processing that can bring about a reduction in irrelevant or noisy data. The study proposed a data extraction algorithm that reduces the data log to almost 80% after removing the irrelevant data. Also, the algorithm proposed in [6] was reported to be efficient in the deletion of redundancy. The study sets a time tolerance threshold and the algorithm proposed in the study for data cleaning was based on some constraints, a better and accurate result was reported.

Redundancy in a dataset may affect the efficiency of models that is created using such data, and in order to avoid disruption in communication that might be caused as a result of redundancy, study in [7], deployed a data cleaning system. This was done by customizing a rate definition to improve the quality of data. Cleaning of data using hybrid approach is the objective of the study proposed in [8]. The study proposed an algorithm that combines a number of techniques to achieve better performance.

The inconsistence in data may also be due to the presence of several duplications. The use of The study in [9], focused on duplicate cleaning to achieve the desires quality data. An algorithm proposed in [10], which filters out clean data from web log files was found to improve the quality of data. Similarly, the study in [11] used statistical and analytical techniques to detect quality problems.

Automatic cleaning and linking of historical census data using the household information was proposed in [12]. Group linkage technique was the method used and the system was reported to have improved the removal of outliers. Although, manual removal of outliers may sometime yield acceptable results, this can only happen when the size of the dataset is very small. Also, to improve the quality of data, study in [13] uses extract transform and load model, the approach reported an efficient data cleaning framework.

Furthermore, the study in [14] proposed an automated detection of outliers in real-world data, the study models the human perception of exceptional values using the concept of fuzzy set theory. The Replicator neural networks was proposed in [15], the study construct models that was used to develop score for outlyingness and provide a measure that promptly unveils outliers in the dataset.

3. MATERIAL AND METHODS

3.1 The Unclean Dataset

The proposed algorithm is represented in this section. In order to implement the algorithm and compare its effectiveness with other technique, an unclean dataset was retrieved from an online open repository [16]. The downloaded file presents health nutrition population data by wealth quintile since 1970s to present. The data have eight attributes and consist of a set of data that comprised of numeric data and several missing values. Some outliers that comprised of alphabets and special characters were added into the dataset in order to populate its inconsistencies. This also helps to determine the efficacy of the proposed algorithm.

3.2 The Proposed Algorithm

As shown in Figure 1, the proposed algorithm is designed to parse through each data contained in the unclean dataset. The essence of the passing was to identify the valid, non-valid or missing data. There are several opinions on how the problem of missing value should be addressed. But the popular opinion as reported in [17], is to compute a standard value to replace the missing value. In view of this, the mean values of the field containing the missing values are used to compute the required mean to effect their replacement. Among the list of measure of central tendency, Mean is the choice here because, its computation is an approximation of the available data in a field.

It should be noted that, when the detected outliers such as alphabets or special characters are identified and removed, this would create more missing values. All the missing values are replaced as earlier explained. The required mean value is computed based on the equation represented in Equation 1.

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (1)$$

where $\overline{X}$ is the mean, x_i represents the set of values in a field and N denotes the size of the data.

Begin
9. *INPUT the unclean dataset*
10. *PARSE through all the data in a field*
11. *DETECT if the data is valid, invalid or missing*
12. *new_data = MEAN (valid data in a field)*
13. *IF the data is valid THEN retain the data*
14. *IF the data is missing or invalid THEN replace the data with the new_data computed*
15. *REPEAT steps 2 – 6 for all the available fields*
16. *DISPLAY the clean dataset*

End.

Figure 1. The proposed algorithm for detection and removal of outliers

3.3 Implementation Of The Proposed Algorithm

The proposed algorithm is implemented using JAVA. Java is a programming language designed to have as few implementation dependency as possible. It is intended to enables the programmers to write their code once and run anywhere. All the eight fields in the unclean dataset explored in this research were searched to detect the valid, invalid and the missing values.

The removal of outliers increased the missing values in the unclean dataset. But after the inconsistencies in the noisy dataset have been detected and removed, then, the entire missing values are replaced with a standard value. As shown in the algorithm, the valid data in a field is used to compute the mean value, this value is then used to replace all the missing data in that field. This is one of the popular view of replacing missing value as reported in [17]. The process continues in other field until all the fields have been covered. The complete clean numeric dataset is then displayed.

3.4 The Use Of Clustering Technique

Clustering technique is one of the methods identified in [4], for the removal of outliers in a dataset. Clustering uses unsupervised classification technique as there are no predefined classes. The objects are grouped together by considering their degree of similarities. The major fundamental of clustering technique can be divided into: partitioning method, the hierarchical method, density-based and grid-based [18].

The use of clustering algorithm performs well in the grouping of similar and dissimilar data, however, it is not designed to fill up the missing values in the original dataset. Also, it is only capable of detecting and grouping the data based on their similarity features, in the process, outliers are distinctly clustered. Studies in [19, 20], have shown that, the clustering techniques only detect outliers in dataset. The issue of missing value is left unresolved. Some of the popular clustering methods used for outlier detection include DBSCAN and BIRCH.

4. RESULTS AND DISCUSSION

The result of implementing the proposed algorithm is represented in this section. The proposed approach implements the popular view of addressing the problem of missing values. One of the popular view is to compute the missing value using the mean of the valid data in the class of individual attribute, this is also the position in [17]. The proposed algorithm, therefore, replaces the missing values and also removes the noisy data as shown in the excerpts of the result generated, (see Table 1).

In order to validate the proposed approach, its effectiveness in the removal of inconsistencies in the dataset is further compared to the use of clustering method for the same task. The clustering algorithm is well reported in the literature, the design of the algorithm focuses on grouping of data. Cluster analysis is one of the exploratory technique that can be used to divide different objects into groups, called clusters; such that, the level of association that exist between the two

objects is maximised if they belong to the same cluster and minimised if otherwise. The comparisons of the proposed algorithm with the clustering technique in the removal of outliers are based on how each algorithm addresses the following in an unclean numeric dataset:

Outlier detection: The proposed algorithm and the clustering methods are both capable of detecting outliers in dataset. The clustering method uses the concept of similarity feature to partition similar and dissimilar data, while the proposed approach parsed through the dataset to detect the missing, valid or invalid data. The data identified as invalid in the present context are data that are not numeric, and they are promptly removed.

Removal of outliers: It is one thing to detect outliers, removing the outlier detected requires further tasks. The proposed algorithm detects and removes outliers in an unclean dataset. Clustering approach is capable of moving the detected outliers into a distinct cluster, this separates it from other data, and however, it is still within the neighbourhood of the entire dataset. But, valid data and outliers are kept in different clusters or clearly separated.

Grouping of outliers: The use of clustering technique, group data based on their similarities, therefore, valid numeric data can easily be separated from the outliers since they don't have anything in common. Each group is otherwise referred to as cluster. The proposed algorithm does not group data, but instead, detect the outliers and delete them. Subsequently, the blank spaces left behind are filled up using a standard value computed.

Replacement of missing value: This metric differentiates the proposed technique from the clustering methods. The clustering methods do not replace missing values as it was not specifically designed to do so. In fact, most of the clustering methods does not tolerate a dataset with missing value, but if zero is used to replace such missing data, then, clustering algorithm would tolerate such dataset. However, the proposed algorithm detects, removes and replaces missing values in a dataset as revealed in the excerpts result of implementing the algorithm represented in the Table 1.

Table 1. The noise and noise-free dataset

A. Noisy dataset

v1	v2	v3	v4	v5	v6	v7	v8
1970				1			1
1971				1	1		2
1972		1		x			1
1973					1		
1974	1		#		1		2
1975		1				1	2
1976				2			2
1977		2	$	1			3
1978		1		1			2
1979		2			1	1	4
1980		2				2	4
1981		y			1		1
1982	1			1			2
1983	2	1		1	2	1	7
1984	3	1		1	&	1	6
1985	3	2			2		7
1986	5	2		2		2	11
1987	4		a	1	1		6
1988	6			1		3	10
1989	8	3	1	1	1	1	15
1990	8			5	3	2	18
1991	9	2		6	4	5	26
1992	11	2	2	5	2	3	25

B. Clean dataset

V1	V2	V3	V4	V5	V6	V7	V8
1970	10.11	3.22	3.98	1	1.39	2.91	1
1971	10.11	3.22	3.98	1	1	2.91	2
1972	10.11	1	3.98	5.35	1.39	2.91	1
1973	10.11	3.22	3.98	5.35	1	2.91	26.09
1974	1	3.22	3.98	5.35	1	2.91	2
1975	10.11	1	3.98	5.35	1.39	1	2
1976	10.11	3.22	3.98	2	1.39	2.91	2
1977	10.11	2	3.98	1	1.39	2.91	3
1978	10.11	1	3.98	1	1.39	2.91	2
1989	10.11	2	3.98	5.35	1	1	4
1980	10.11	2	3.98	5.35	1.39	2	4
1981	10.11	3.22	3.98	5.35	1.39	1	1
1982	1	3.22	3.98	1	1.39	2.91	2
193	2	1	3.98	1	2	1	7
1984	3	1	3.98	1	1.39	1	6
1985	3	2	3.98	5.35	1.39	2	7
1986	5	2	3.98	2	1.39	2	11
1987	4	3.22	3.98	1	1	2.91	6
1988	6	3.22	3.98	1	1.39	3	10
1989	8	3	1	1	1	1	15
1990	8	3.22	3.98	5	3	2	18
1991	9	2	3.98	6	4	5	26
1992	11	2	2	5	2	3	25

Table 2. Comparison of the proposed cleaning algorithm with the use of clustering algorithm

Approaches	Metrics			
	Detection of outliers	Removal of Outliers	Grouping of outliers	Replacement of missing values
Proposed Method	✔	✔	x	✔
Clustering Method	✔	x	✔	x

Note : ✔ supported x not supported

5. CONCLUSIONS

This paper has shown how inconsistencies in the numeric dataset can be removed through the implementation of the proposed algorithm. The paper focuses on an aspect of data pre-processing, data cleaning. The study is aimed at exploring the unclean dataset, with a view to putting them in a suitable format for exploratory purposes. Some approaches reported in the literature for performing similar task include: clustering, associative, pattern-based and statistical methods. The effectiveness of the proposed algorithm is compared to the use of one of the cleaning approaches listed, clustering, on the set of unclean data. The comparisons were based on what the algorithm of clustering is capable of doing and what is not designed to do as reported in the literature. Findings from the comparison of the two approaches show that, the proposed approach is capable of producing a more efficient and better accurate clean dataset. The Table 1 that comprised of the excerpts of the original and cleaned dataset also confirm the effectiveness of the proposed algorithm. The proposed approach is found to be suitable for removing the anomalies or noise in a numeric dataset irrespective of the size of the records been explored.

References

[1] M. Kantardzic, *DATA MINING: Concepts, Models, Methods and Algorithms*, 2nd ed. New Jersey: IEEE Press, John Wiley & Sons, Inc., 2011.

[2] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence,* vol. 17, pp. 375-381, 2003.

[3] X. Chu, I. F. Ilyas, and P. Papotti, "Holistic data cleaning: Putting violations into context," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, 2013, pp. 458-469.

[4] I. Ahmed and A. Aziz, "Dynamic approach for data scrubbing process," *International Journal on Computer Science and Engineering,* vol. 2, pp. 416-423, 2010.

[5] M. Srivastava, R. Garg, and P. Mishra, "Analysis of Data Extraction and Data Cleaning in Web Usage Mining," in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, 2015, p. 13.

[6] K. Hu, L. Li, C. Hu, J. Xie, and Z. Lu, "A dynamic path data cleaning algorithm based on constraints for RFID data cleaning," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on*, 2014, pp. 537-541.

[7] A. Ebaid, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, J.-A. Quiane-Ruiz, N. Tang*, et al.*, "NADEEF: A generalized data cleaning system," *Proceedings of the VLDB Endowment,* vol. 6, pp. 1218-1221, 2013.

[8] A. Paul, V. Ganesan, J. S. Challa, and Y. Sharma, "HADCLEAN: A hybrid approach to data cleaning in data warehouses," in *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*, 2012, pp. 136-142.

[9] L. He, Z. Zhang, Y. Tan, and M. Liao, "An efficient data cleaning algorithm based on attributes selection," in *Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference on*, 2011, pp. 375-379.

[10] S. Anand and R. Rani Aggarwal, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions," *International Journal of Computer Applications,* vol. 48, pp. 13-18, 2012.

[11] L. Berti-Equille, T. Dasu, and D. Srivastava, "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," in *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, 2011, pp. 733-744.

[12] Z. Fu, P. Christen, and M. Boot, "Automatic cleaning and linking of historical census data using household information," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 413-420.

[13] H. Hj Mohamed, T. L. Kheng, C. Collin, and O. S. Lee, "E-Clean: A Data Cleaning Framework for Patient Data," in *Informatics and Computational Intelligence (ICI), 2011 First International Conference on*, 2011, pp. 63-68.

[14] M. Last and A. Kandel, "Automated detection of outliers in real-world data," in *Proceedings of the second* ***international conference on intelligent technologies*****, 2001, pp. 292-301.**

[15] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in *Data warehousing and knowledge discovery*, ed: Springer, 2002, pp. 170-180.

[16] Databank. (2014). *Health Nutrition and Population Statistics by Wealth Quintile, World Bank Group*. Available: http://data.worldbank.org/data-catalog/HNPquintile

[17] S. Tuffery, *Data Mining and Statistics for Decision Making*. USA: John Wiley & Sons Ltd, 2011.

[18] J. Han, M. Kamber, and J. Pei, *Data mining, southeast asia edition: Concepts and techniques*: Morgan kaufmann, 2012.

[19] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, pp. 226-231.

[20] A. Loureiro, L. Torgo, and C. Soares, "Outlier detection using clustering methods: a data cleaning application," in *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany*, 2004.