International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229212

An Applied Mean Substitutions Technique for Detection of Anomalous Value in Data Mining

Dr. Darshanaben Dipakkumar Pandya[1], Dr. Abhijeetsinh Jadeja[2], Dr. Sheshang D. Degadwala[3]

[1]Assistant Professor, Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India

[2]Principal(I/C), Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India

[3]Head of Computer Department, Sigma Institute of Engineering, Vadodara, Gujarat, India

Article Info
Volume 9, Issue 2
Page Number : 103-108

Publication Issue
March-April-2022

Article History
Accepted : 01 April 2022
Published : 05 April 2022

ABSTRACT

In the numerical value database, inliers in a database are subset of observations adequately small enough compared to the rest of the observations, which appears to be inconsistent with the remaining data set. They are the result of instant failures or early failures, experienced in many life-test experiments. The problem is how to handle Inliers in a dataset, and how to evaluate the Inliers. This paper describes a revolutionary of approach that uses Inliers detection as a pre-processing step to detect the Inliers and then applies Mean Substitution technique algorithm, hence to analyze the effects of the Inliers on the analysis of dataset.

Keywords: Data Mining, Attribute, Inliers Detection Approach Algorithm, Mean Substitution Technique Algorithm

I. INTRODUCTION

An anomalous value in database is solitary of the principle problems featured in data analysis and in the prediction. The belongings of these anomalous values are highly reflected on the final results. Our chief goal is to achieve the final result without error in the consolidated form, which is use to take decisions. Now let us consider the following example as a natural occurrence of a physical phenomenon: 0, 0, 0, 0, 0.01, 0.03, 0.08, 1.50, 1.96, 1.21, 1.75, 2.53, 3.90 and 4.10. Here, the first four observations may be treated as instantaneous failures, next three observations may be treated as early failures and other observations may be treated as coming from any failure time distribution.

In this study, a method of Inliers detection is introduced and discussed which provides an approach to treat anomalous values. This step treats the anomalous block of values from a real-world imbalanced database.

II. Background on Anomalous Data

In this study, a statistical method is discussed which provides an approach to find out pattern to discover anomalous values from a real imbalanced database with massive anomalous values. Therefore, the objective of this method is to discover the best fitted value for the anomalous value and select records completely by removing Inliers.

The function of statistical methods has gained stuff in exploring evaluation and calculation techniques. Lee, J., & Wonpil, Y[1] are the authors who have introduced Concurrent Tracking of Inliers and Outliers. Winkler, W[2] investigated Problems with inliers. Muralidharan K. and Arti M [3] investigated analysis of instantaneous and early failures in Pareto distribution. Muralidharan, K. and B. K. Kale [4] are the authors who have introduced Inliers detection using Schawartz information criterion, K. Muralidharan, Arti. Khabia[5] introduced Inliers proness in normal distribution. K. Muralidharan [6] are the scientists who invested theory of inliers modeling and applications. Winklers, W. E [7] are the authors who have introduced Problems with inliers. The objective of proposed study is to determine the statistical technique which may be significant in the handling of anomalous attribute values.

III. Inliers Analysis

An inlier's is a data observation that lies in the interior of a data set and is unusual or in error. Because inliers are difficult to distinguish from the other data values, they are sometimes difficult to find and -if they are in error to correct. The descriptive analysis applied on data. Following results, shown in table 1 with inliers and in table 2 without inliers, whereas that of analysis of recovered data is shown in table 3.

Table 1. Descriptive analysis OF PRE-MONSOON the data based on table-4 (with Inliers) Descriptive Statistics (with inliers)

RAIN (IN MILIMETERE)	SUM	MEAN	MEDIAN	MODE	S.D	C.V
Pre-Monsoon	1563.66	31.27	35.80	3.000	19.9	0.64
Monsoon	2019.3	40.39	36.85	3.00	30.6	0.76
Post-Monsoon	553.5	11.07	8.00	3.00	10.2	0.93

Table 2. Descriptive analysis of MONSOON the data based on table-4 (without Inliers) Descriptive Statistics (without Inliers)

RAIN (IN MILIMETERE)	SUM	MEAN	MEDIAN	MODE	S.D	C.V
Pre-Monsoon	1543.8	37.7	37.7	54.00	20.5	0.55
Monsoon	1994.3	51.1	46.8	80.4	31.3	0.61
Post-Monsoon	530.3	13.3	8.9	6.4	10.7	0.80

Table 3. Descriptive analysis of MONSOON the data based on table-4 (Recovered)

Descriptive Statistics (Recovered)

RAIN (IN MILIMETERE)	SUM	MEAN	MEDIAN	MODE	S.D	C.V
Pre-Monsoon	1883.1	37.7	37.7	37.7	14.4	0.38
Monsoon	2556.4	51.1	51.1	51.1	22.8	0.45
Post-Monsoon	663.3	13.3	9.8	13.3	9.2	0.70

The below table-4 shows Inliers Detection approach of the dataset with Inliers and treatment by removing it from database and recovering missing values use Mean Substitution technique for data recovering.

IV. PROPOSED APPROACH

As reviewed several different ways of detecting Inliers here propose a method which is a combination of different approaches, statistical and data mining. Firstly apply Inlier's detection using Inliers Detection approach algorithm to group the data into parts for discovering Inliers and removing it from dataset and then Mean Substitution algorithm for recovering the missing values from the dataset. The below figure shows the overall idea.

System Architecture

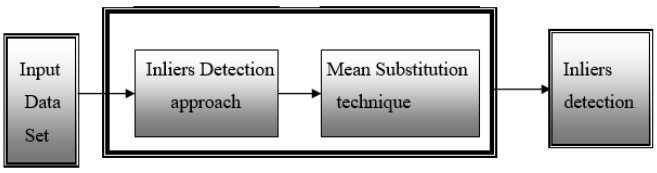

Fig. 2. System Architecture

4.1 Inliers Detection approach algorithm

The proposed method is based on finding inliers value from the data set by the Inliers Detection approach method. In general, this method is search of Inlier's value which is very close to the true mean of the attribute. If found Inliers then remove the data entry having Inliers permanently from the data set depending upon the Inliers detection criteria.

Introduction: Given an array R of size N, this procedure finds the elements of having Inlier's values. The variable Min_Index shows the minimum value for Inliers finding in data set. Here we take Min_Range variable which indicate size of minimum for finding Inliers in a data set respectively. The variable I is used to index elements from 1 to N in a given pass.

Following are the steps of the algorithm in detail:

Step 1: Select a dataset on which Inlier's detection is to be performed from the database.

Step 2: Initialization of variables.

Min_Index ← 05.

Step 3: Create a loop for N passes

Repeat through step 5 for I = 1, 2… N.

Step 4: Make a pass and obtain element with Inlier's value.

If R [I] < Min_Index

Write 'Inliers found in the data set '

then R [I] = NULL // Assigning NULL value to array.

Write 'Inliers Removed from the data set '.

else

Write 'Inliers not found in the data set'.

Step 6: finished.

4.2 Mean Substitution technique algorithm

The intended method is based on replacing missing attribute values by the Mean Substitution technique method. This method is very much helpful for numerical attributes. In general, this method is search of missing values and after searching its value is replaced by the mean of the attribute and closest to

the value of just preceding and succeeding value of the missing values.

Introduction: Given an array R of size N, this procedure replaces the missing values with the recovered data from the Inlier's data set. The variable I is used to index elements from 1 to N in a given data. Following are the steps of the algorithm in detail:

Step 1: Select a dataset on which Missing values recovery is to be performed from the database.

Step 2: Initialize

Mean ← NULL.

I ← NULL.

Step 3: Determine the mean from the data using

Mean = $\frac{X1 + X2 + X3 \ldots\ldots Xn}{N}$ Or Mean = $\frac{Exi}{N}$

Step 4: Create a loop for N passes

Repeat through step 8 for I = 1, 2… N.

Step 5: Perform Missing value Recovery Process from Inliers database.

do

If (R [I] = = NULL)

then

R [I] = Mean // Estimated value

Step 6: Make iterations of each pass.

I = I +1. // Iterations

Step 7: Iteration is to be performed till condition is satisfied.

Repeat until (I > = N)

Step 8: Finished.

Stop.

V. Discussion of Results

Measure of central tendency (mean): Table-1 shows the seasonal distribution of average rainfall in different districts in Gujarat from 1955-2014(Rain fall in millimeter) dataset of average rainfall from analysis by season type Pre-monsoon, monsoon, post-monsoon. The mean of average rainfall in different districts in Pre-monsoon, monsoon and post-monsoon are 31.27, 40.39 and 11.07 respectively. After missing values at the extremes, the mean calculated from incomplete data sets are 37.7 for Pre-monsoon, 51.1 for monsoon and 13.3 for post-monsoon.

The proposed mean substitution method is applied on the data sets of Table 1 to fill up the missing values. It is observed that mean values of Pre-monsoon, monsoon and post-monsoon are 37.7, 51.1 and 13.3 respectively. It is considerable that the mean values obtained after replacing the missing values by the proposed approach very same as the actual mean as given.

Median and Mode: From the analysis of result of Median and Mode it is found that after estimation of missing values, the values of Median and Mode obtained are close to the Median and Mode of standard dataset. On the basis of result there can be said that proposed algorithm is appropriate for Inliers finding and detection of Inliers also recovery of the data.

Standard Deviation: From the analysis of result of standard deviation it is found that after estimation of missing values, the values of standard deviation obtained are close to the standard deviation of standard dataset. On the basis of result there can be said that proposed algorithm is appropriate for Inliers finding and detection of Inliers also recovery of the data.

Coefficient of Variation: From the analysis of result of co-efficient of variation (CV) it is found that, after estimation of missing values, the values of co-efficient of variation is very near , which shows that the series is uniform now. It is observed that recovered Standard deviation values are varying close to outliers removed dataset.

VI. Experimental Results

There can be a hypothetical data which has been made by introducing some Inliers values in the well known rainfall data. The above table 4 shows Mean Substitution technique of the dataset with Inliers. Now must delete the Inliers entry and save both the dataset i.e. with Inliers entry and without Inliers

entry and run further the Inliers detection Approach algorithm and Mean Substitution technique approach to do the analysis of the data and calculate the sum of points to the value in each case.

VII. Conclusion

The conclusion lies in the fact that Inliers are usually the unwanted entries which always affects the data in one or the other form and misreports the distribution of the data. Sometimes it becomes necessary to keep even the Inliers entries because they play an important role in the data but in our case achieving and our main objective is to discovering Inliers entries and i.e. to delete the Inliers entries from database. Proposed approach provides proper consolidated report using data relative attributes of the database.

VIII. REFERENCES

[1]. Lee, J., & Wonpil, Y. (2014). Concurrent Tracking of Inliers and Outliers.

[2]. Winkler, W. (1998), Problems with inliers. Retrieved October 5, 2015.

[3]. Muralidharan K. and Arti M. investigation of instantaneous and early failures in Pareto distribution, Journal of statistical theory and Applications, Vol. 7, 2008, pp. 187–204.

[4]. Muralidharan, K. and B. K. Kale Inlier detection using Schawartz information criterion. J. Reliability and Stat. Studies, Vol. 1(1), 2008, pp.1–5.

[5]. K. Muralidharan, Arti. Khabia, Inliers proness in normal distribution, (Vol.8) 2013, March.

[6]. K. Muralidharan, theory of inliers modeling and applications, University of Bedfordshire, 2011.

[7]. Winkler, W. E. (1997). Problems with inliers. Paper presented at the European Conference of Statisticians, Prague. last accessed 28 May 2014.

Table 4: Mean Substitution technique approach of the dataset with and without Inliers.

Dataset of seasonal distribution of average rainfall in different districts in Gujarat from 1955-2014(Rainfall in millimeter).

Standard Data					Inliers Results in data			Inliers Removed and Missing values obtained			Recovered Values		
SN	YEAR	Pre-Monsoon	Monsoon	Post-monsoon	Pre-Monsoon	Monsoon	Post-Monsoon	Pre-Monsoon	Monsoon	Post-Monsoon	Pre-Monsoon	Monsoon	Post-Monsoon
1	1955	49.9	77.5	9.2	FALSE	FALSE	FALSE	49.9	77.5	9.2	49.9	77.5	9.2
2	1956	31.6	3	5.3	FALSE	TRUE	FALSE	31.6	___	5.3	31.6	51.1	5.3
3	1957	0	79.9	5.2	TRUE	FALSE	FALSE	___	79.9	5.2	37.7	79.9	5.2
4	1958	36.8	83.2	0	FALSE	FALSE	TRUE	36.8	83.2	___	36.8	83.2	13.3
5	1959	33.6	0	5.4	FALSE	TRUE	FALSE	33.6	___	5.4	33.6	51.1	5.4
6	1960	63.6	77.8	12.3	FALSE	FALSE	FALSE	63.6	77.8	12.3	63.6	77.8	12.3
7	1961	54	80.4	11.4	FALSE	FALSE	FALSE	54	80.4	11.4	54	80.4	11.4
8	1962	43.7	81.2	3	FALSE	FALSE	TRUE	43.7	81.2	___	43.7	81.2	13.3
9	1963	44.1	3	8.8	FALSE	TRUE	FALSE	44.1	___	8.8	44.1	51.1	8.8
10	1964	37	80.4	8.3	FALSE	FALSE	FALSE	37	80.4	8.3	37	80.4	8.3
11	1965	3	82.3	6.1	TRUE	FALSE	FALSE	___	82.3	6.1	37.7	82.3	6.1
12	1966	35.4	78	8.8	FALSE	FALSE	FALSE	35.4	78	8.8	35.4	78	8.8
13	1967	31.3	3	7.6	FALSE	TRUE	FALSE	31.3	___	7.6	31.3	51.1	7.6
14	1968	44.2	80.1	3	FALSE	FALSE	TRUE	44.2	80.1	___	44.2	80.1	13.3

15	1969	37.3	83.1	6		FALSE	FALSE	FALSE		37.3	83.1	6		37.3	83.1	6
16	1970	44.3	84.4	6.3		FALSE	FALSE	FALSE		44.3	84.4	6.3		44.3	84.4	6.3
17	1971	33.7	3	36.1		FALSE	TRUE	FALSE		33.7	___	36.1		33.7	51.1	36.1
18	1972	6.5	65.7	1		FALSE	FALSE	TRUE		6.5	65.7	___		6.5	65.7	13.3
19	1973	1	11.8	6.4		TRUE	FALSE	FALSE		___	11.8	6.4		37.7	11.8	6.4
20	1974	64.9	7.1	7.8		FALSE	FALSE	FALSE		64.9	7.1	7.8		64.9	7.1	7.8
21	1975	44.7	6	6.4		FALSE	FALSE	FALSE		44.7	6	6.4		44.7	6	6.4
22	1976	55	21	6.4		FALSE	FALSE	FALSE		55	21	6.4		55	21	6.4
23	1977	37.7	3	8.4		FALSE	TRUE	FALSE		37.7	___	8.4		37.7	51.1	8.4
24	1978	65	69.4	2		FALSE	FALSE	TRUE		65	69.4	___		65	69.4	13.3
25	1979	22	36.7	6.7		FALSE	FALSE	FALSE		22	36.7	6.7		22	36.7	6.7
26	1980	3	36.4	7.4		TRUE	FALSE	FALSE		___	36.4	7.4		37.7	36.4	7.4
27	1981	54	37.8	8.9		FALSE	FALSE	FALSE		54	37.8	8.9		54	37.8	8.9
28	1982	62.1	61	9.7		FALSE	FALSE	FALSE		62.1	61	9.7		62.1	61	9.7
29	1983	12.4	37	3.6		FALSE	FALSE	TRUE		12.4	37	___		12.4	37	13.3
30	1984	36.4	34	15.1		FALSE	FALSE	FALSE		36.4	34	15.1		36.4	34	15.1
31	1985	14.3	2	36.7		FALSE	TRUE	FALSE		14.3	___	36.7		14.3	51.1	36.7
32	1986	13.5	33	16.8		FALSE	FALSE	FALSE		13.5	33	16.8		13.5	33	16.8
33	1987	3	37.6	25		TRUE	FALSE	FALSE		___	37.6	25		37.7	37.6	25
34	1988	56.4	23.1	21.3		FALSE	FALSE	FALSE		56.4	23.1	21.3		56.4	23.1	21.3
35	1989	41.3	33.4	3.6		FALSE	FALSE	TRUE		41.3	33.4	___		41.3	33.4	13.3
36	1990	7.8	3	6.9		FALSE	TRUE	FALSE		7.8	___	6.9		7.8	51.1	6.9
37	1991	8.6	34.7	9.7		FALSE	FALSE	FALSE		8.6	34.7	9.7		8.6	34.7	9.7
38	1992	0.36	44.3	10.3		TRUE	FALSE	FALSE		___	44.3	10.3		37.7	44.3	10.3
39	1993	24.6	46.8	9.8		FALSE	FALSE	FALSE		24.6	46.8	9.8		24.6	46.8	9.8
40	1994	34.4	2	6.3		FALSE	TRUE	FALSE		34.4	___	6.3		34.4	51.1	6.3
41	1995	2	23.9	3		TRUE	FALSE	TRUE		___	23.9	___		37.7	23.9	13.3
42	1996	45.9	46.1	13.4		FALSE	FALSE	FALSE		45.9	46.1	13.4		45.9	46.1	13.4
43	1997	46	63	8.2		FALSE	FALSE	FALSE		46	63	8.2		46	63	8.2
44	1998	50.3	0	44		FALSE	TRUE	FALSE		50.3	___	44		50.3	51.1	44
45	1999	4	48	3		TRUE	FALSE	TRUE		___	48	___		37.7	48	13.3
46	2000	36.2	65	23.7		FALSE	FALSE	FALSE		36.2	65	23.7		36.2	65	23.7
47	2001	32	3	36.1		FALSE	TRUE	FALSE		32	___	36.1		32	51.1	36.1
48	2002	36.7	16.4	32.8		FALSE	FALSE	FALSE		36.7	16.4	32.8		36.7	16.4	32.8
49	2003	3.5	78	1		TRUE	FALSE	TRUE		___	78	___		37.7	78	13.3
50	2004	14.6	8.8	9.3		FALSE	FALSE	FALSE		14.6	8.8	9.3		14.6	8.8	9.3

SUM	1563.66	2019.3	553.5	1543.8	1994.3	530.3	1883.1	2556.4	663.3
MEAN	31.27	40.39	11.07	37.7	51.1	13.3	37.7	51.1	13.3
MEDIAN	35.80	36.85	8.00	37.0	46.8	8.9	37.7	51.1	9.8
MODE	3.00	3.00	3.00	54.0	80.4	6.4	37.7	51.1	13.3
SD	19.9	30.6	10.2	20.5	31.3	10.7	14.4	22.8	9.2
CV	0.64	0.76	0.93	0.55	0.61	0.80	0.38	0.45	0.70

Cite this article as :
Dr. Darshanaben Dipakkumar Pandya, Dr. Abhijeetsinh Jadeja, Dr. Sheshang D. Degadwala, "An Applied Mean Substitutions Technique for Detection of Anomalous Value in Data Mining ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 2, pp. 103-108, March-April 2022. Available at doi : https://doi.org/10.32628/IJSRST229212
Journal URL : https://ijsrst.com/IJSRST229212