International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2285201

Closest fit Approach for Atypical Value Revealing and Deciles Range Anomaly Detection Method for Recovering Misplaced value in Data Mining

Dr. Darshanaben Dipakkumar Pandya[1], Bhumika Kumarbhai Modi[2], Nidhi S. Bhavsar[3]
[1]Assistant Professor, Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India
[2]Assistant Professor, Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India
[3]Research Scholar, Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan, India

Article Info

Publication Issue :
Volume 8, Issue 5
September-October-2022

Page Number : 217-222

Article History
Accepted: 17 Oct 2022
Published: 18 Oct 2022

ABSTRACT

In identifying anomalous database values, it is currently a very active research area in the mining community. The task of identifying anomalous values is to find a small group of exceptional data objects compared to the rest of the large amount of data. The discovery of anomalous values in a group of models is an extremely recognized difficulty in the field of data mining. An outlier is a prototype that is not related to the rest of the patterns in the data set. The proposed method for searching for outliers uses an anomalous detection approach. The purpose of the approach is to find anomalous values first based on the criteria of the condition.We use the information criterion and approach named Outlier Detection to remove the outliers from the dataset and apply deciles range anomaly detection algorithm for Recovery algorithm to recover missing data from the database.

Keywords: Data Mining, Anomalous Values, Outlier Detection Approach, Deciles Range Anomaly Detection Algorithm, Recovery.

I. INTRODUCTION

An outlier is defined as a data point that is very different from the rest of the data. The anomalous detection technique finds applications in financial applications, credit card fraud, intrusion detection in the network and marketing. The atypical result is the search for objects in the database that do not follow the appropriate laws for most of the data. The recognition of an object as anomalous is influenced by several factors, many of which are interesting for the practical applications available.

2. Background on atypical Data Outliers and absent values:

In this study, we discuss a method that provides an approach to uncover patterns to uncover abnormal values from a true unbalanced database with massive anomalous values. Therefore, the goal of this method is to find out the correct best value for the anomalous value and select the records completely by eliminating the outliers and retrieving the missing values.

C. Aggarwal and P. Yu[1] are the persons who have introduced Outlier Detection for High Dimensional Data in various ways. Z. He, X. Xu and S. Deng [2] are the persons who have discovering Cluster based

Local Outliers and various techniques. S. Lin and D. Brown [3] recommended An Outlier-based Data Association Method using various techniques. S. Ramaswamy, R. Rastogi and K. Shim [4] give a fabulous and Efficient Algorithms for Mining Outliers from Large Data Sets. Lu C., Chen D, Kou Y [5] explored about algorithms for spatial outlier detection using various methods. Kim, J.O., and Curry, J [6] discussed about the treatment of missing data in multivariate analysis. Qin, Y.S. [7] convoluted Semi-parametric optimization for missing data imputation, Applied Intelligence.

II. Outlier Analysis

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In the analysis of anomalous data mining values can be done with different methods. The proposed method is based on the replacement of the values of atypical attributes discovered and therefore on the elimination of data that have anomalous parental values from the data set. This method is very useful for numeric attributes. In general, this method is a search for values very close to the real mean of the attribute.

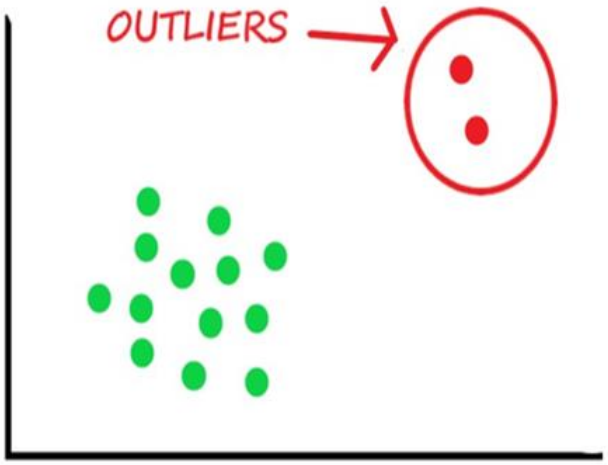

Fig.1. Diagram of outlier detection example in data set

III. Proposed Approach

Since we have reviewed numerous dissimilar ways of detecting outliers now suggest a method which is a grouping of different approaches, statistical and data mining. Initially apply outlier's finding using Outlier Detection algorithm to group the data into parts for discovering outliers and removing it from dataset and then Missing Block Odd size Recovery algorithm for recovering the missing values from the dataset.

4.1 Outlier Detection algorithm

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

Introduction: Given an array K of size N, this procedure finds the elements of having outlier's values. The variable Outlier_Index shows the maximum value for outliers finding in data set. Here we take Max_data variables which indicate size of maximum for finding outliers 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 outlier's detection is to be performed from the database.

Step 2: Initialize

Max_data ← 2000.

Step 3: Create a loop for N passes

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

Step 4: Initialize Outlier_Index to size of limitation of each pass.

Outlier_Index ← Max_data.

Step 5: Make a pass and obtain element with outlier's value.

If K [I] > = Outlier_Index

Write 'Outliers found in the data set '

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

Write 'Outliers Removed from the data set '.
else
Write 'Outliers not found in the data set'.
Step 6: finished.

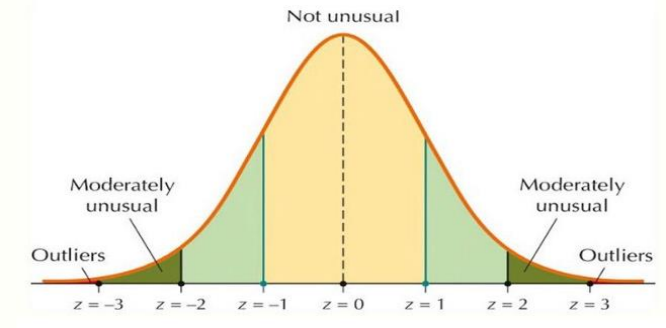

Fig.2. Diagram of detection of outliers with Z-Score in data set

4.2. Deciles range anomaly detection method algorithm

The proposed method is based on replacing absent values by the deciles range anomaly Approach method. This method is very much useful for numerical attributes. In general, this method is search of misplaced value which is very close to the true mean of the attribute and closest to the value of just preceding and succeeding value of the missing values.

Introduction: Given an array K of size N, this procedure replaces the missing values with the recovered data from the 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 anomalous value detection is to be performed from the database.

Step 2: Determine the 0.7^{th} deciles value using Percentile function and 0.7 values divide by 10 and 9.2^{th} deciles value using Percentile function and 9.2 values divide by 10 of the data given from dataset.

Step 3: Determine the fences. Fences serve as cutoff points for determining anomalous values from the database.

Lower fence A = 0.7^{th} deciles value using (Percentile function and 0.7 values
divide by 10)

Upper fence B = 9.2th deciles value using (Percentile function and 9.2 values
divide by 10)

Step 4: If a data value is less than the lower fence A then it is considered as inliers or greater than the upper fence B, it is considered an outlier in dataset.

Step 5: Then run the algorithm, for the dataset with and without the data having anomalous data, using replicates in order to select proper centroids so as to overcome the problem of local minima from the dataset.

Step 6: If the step 5 conditions is true, then remove the data entry having anomalous data permanently from the dataset.

Step 7: finished.

Stop.

IV. DISCUSSION OF RESULTS

Table-1 given in appendix shows the world wide emission of carbon dioxide (CO2) from the consumption of Coal, Oil and Natural Gas respectively for the years 1960 to 2009. The mean emission of carbon dioxide (CO2) due to Coal, Oil and Natural Gas are 1449.42, 939.98 and 1659.02 respectively. After missing values at the extremes, the mean calculated from incomplete data sets are 1214.37 for Coal, 723.61 for Oil and 1494.34 for Natural Gas. It is observed that mean values of incomplete data sets are lower than the mean values from the standard dataset.

The proposed ratio based approach method is applied on the data sets of Table 1 to fill up the missing values. It is observed that mean values of Coal, Oil and Natural Gas are 1191.02, 751.71 and 1488.52 respectively. It is considerable that the mean values obtained after replacing the missing values by the proposed approach is lower than the actual mean as given.

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 we can say that proposed algorithm is appropriate for outliers finding and detection of outliers 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.

V. CONCLUSION

This paper shows the universal truth that there is no accurate method of Finding outliers and treatment missing attribute values. The proposed approach is an important one for the arithmetical real value having deviation from the mean due to their presence in the attribute. This approach gives proper result for the consolidated report which is generated from the database. The proposed even size block fitting approach is useful for numerical attribute, having minor deviation from the mean. The method is appropriate for the consolidated report, also more appropriate and suitable to small size block missing values in data mining.

VI. REFERENCES

[1]. Darshanaben Dipakkumar Pandya, Dr. Sanjay Gaur, “Inliers Detection and Recovered Missing value in Data Mining”, International Journal of Emerging Technology and Advanced Engineering, Volume 8, Special Issue4, pp.1-6, April 2018.

[2]. C. Aggarwal and P. Yu, “Outlier Detection for High Dimensional Data”. In Proceedings of the International Conference on Management of Data, Volume 30, Issue 2, pages 37 – 46, May 2001.

[3]. Z. He, X. Xu and S. Deng, “Discovering Cluster based Local Outliers”. Pattern Recognition Letters, Volume 24, Issue 9-10, pages 1641 – 1650, June 2003.

[4]. Darshanaben Dipakkumar Pandya, Dr. Sanjay Gaur, Detection of Anomalous value in Data Mining, Kalpa Publications in Engineering, Volume 2, pp.1-6, 2018

[5]. S. Ramaswamy, R. Rastogi and K. Shim, “Efficient Algorithms for Mining Outliers from Large Data Sets”. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Volume 29, Issue 2, pages 427 – 438, May 2000.

[6]. Darshanaben Dipakkumar Pandya, Dr. Sanjay Gaur, “Closest Fit Approach for Pattern Designing to Recovered Anomalous Values in Data Mining”, International Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 308 - 312, 2018

Cite this article as :

Dr. Darshanaben Dipakkumar Pandya, Bhumika Kumarbhai Modi, Nidhi S. Bhavsar, "Closest fit Approach for Atypical Value Revealing and Deciles Range Anomaly Detection Method for Recovering Misplaced value in Data Mining", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 5, pp. 217-222, September-October 2022. Available at doi : https://doi.org/10.32628/CSEIT2285201

Journal URL : https://ijsrcseit.com/CSEIT2285201

Table : 1

Outlier Detection approach of the dataset with outliers and deciles anomaly detection method for Recovery Approach for data recovering.

Global Carbon Dioxide Emissions from Fossil Fuel Burning by Fuel Type, 1960-2009 (In Million Tones of Carbon

Standard Data | **Outliers Results** | **Outliers Removed** | **Recovered Values**

Missing values obtained

S.N	YEAR	COAL	OIL	NATURAL GAS	COAL	OIL	NATURAL GAS	COAL	OIL	NATURAL GAS	COAL	OIL	NATURAL GAS
		Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon		
1	1960	1410	849	1235	FALSE	FALSE	FALSE	1410	849	1235	1410	849	1235
2	1961	2349	904	1254	TRUE	FALSE	FALSE	___	904	1254	1416	904	1254
3	1962	2351	980	1277	TRUE	FALSE	FALSE	___	980	1277	1423	980	1277
4	1963	2396	1052	1300	TRUE	FALSE	FALSE	___	1052	1300	1429	1052	1300
5	1964	1435	1137	2328	FALSE	FALSE	TRUE	1435	1137	___	1435	1137	1328
6	1965	1460	1219	2351	FALSE	FALSE	TRUE	1460	1219	___	1460	1219	1355
7	1966	1478	1323	2380	FALSE	FALSE	TRUE	1478	1323	___	1478	1323	1383
8	1967	1448	1423	1410	FALSE	FALSE	FALSE	1448	1423	1410	1448	1423	1410
9	1968	1448	1551	1446	FALSE	FALSE	FALSE	1448	1551	1446	1448	1551	1446
10	1969	1486	1673	1487	FALSE	FALSE	FALSE	1486	1673	1487	1486	1673	1487
11	1970	1556	1839	1516	FALSE	FALSE	FALSE	1556	1839	1516	1556	1839	1516
12	1971	1559	1946	1554	FALSE	FALSE	FALSE	1559	1946	1554	1559	1946	1554
13	1972	1576	2055	1584	FALSE	TRUE	FALSE	1576	___	1584	1576	1699	1584
14	1973	1581	2240	1608	FALSE	TRUE	FALSE	1581	___	1608	1581	1451	1608
15	1974	1579	2244	1618	FALSE	TRUE	FALSE	1579	___	1618	1579	1204	1618
16	1975	1673	956	1623	FALSE	FALSE	FALSE	1673	956	1623	1673	956	1623
17	1976	1710	564	1650	FALSE	FALSE	FALSE	1710	564	1650	1710	564	1650
18	1977	1766	514	1649	FALSE	FALSE	FALSE	1766	514	1649	1766	514	1649
19	1978	1793	392	1677	FALSE	FALSE	FALSE	1793	392	1677	1793	392	1677
20	1979	887	544	1719	FALSE	FALSE	FALSE	887	544	1719	887	544	1719
21	1980	947	422	1740	FALSE	FALSE	FALSE	947	422	1740	947	422	1740
22	1981	921	289	1756	FALSE	FALSE	FALSE	921	289	1756	921	289	1756
23	1982	2677	196	1746	TRUE	FALSE	FALSE	___	196	1746	1130	196	1746
24	1983	2719	177	1745	TRUE	FALSE	FALSE	___	177	1745	1339	177	1745
25	1984	2740	202	1808	TRUE	FALSE	FALSE	___	202	1808	1547	202	1808
26	1985	1756	182	1836	FALSE	FALSE	FALSE	1756	182	1836	1756	182	1836
27	1986	1746	290	1830	FALSE	FALSE	FALSE	1746	290	1830	1746	290	1830
28	1987	1745	302	1893	FALSE	FALSE	FALSE	1745	302	1893	1745	302	1893

29	1988	1808	408	1936		FALSE	FALSE	FALSE		1808	408	1936		1808	408	1936
30	1989	1836	455	2972		FALSE	FALSE	**TRUE**		1836	455	___		1836	455	**1727**
31	1990	1830	517	2026		FALSE	FALSE	**TRUE**		1830	517	___		1830	517	**1519**
32	1991	1893	627	2069		FALSE	FALSE	**TRUE**		1893	627	___		1893	627	**1310**
33	1992	1936	506	1101		FALSE	FALSE	FALSE		1936	506	1101		1936	506	1101
34	1993	455	537	1119		FALSE	FALSE	FALSE		455	537	1119		455	537	1119
35	1994	517	562	1132		FALSE	**TRUE**	FALSE		517	___	1132		517	**501**	1132
36	1995	627	564	1153		FALSE	**TRUE**	FALSE		627	___	1153		627	**465**	1153
37	1996	506	514	1208		FALSE	**TRUE**	FALSE		506	___	1208		506	**428**	1208
38	1997	537	392	1211		FALSE	FALSE	FALSE		537	392	1211		537	392	1211
39	1998	562	544	1245		FALSE	FALSE	FALSE		562	544	1245		562	544	1245
40	1999	564	422	1272		FALSE	FALSE	FALSE		564	422	1272		564	422	1272
41	2000	2514	289	1291		**TRUE**	FALSE	FALSE		___	289	1291		**529**	289	1291
42	2001	2392	196	1314		**TRUE**	FALSE	FALSE		___	196	1314		**493**	196	1314
43	2002	2544	177	1349		**TRUE**	FALSE	FALSE		___	177	1349		**458**	177	1349
44	2003	422	202	1399		FALSE	FALSE	FALSE		422	202	1399		422	202	1399
45	2004	289	849	1436		FALSE	FALSE	FALSE		289	849	1436		289	849	1436
46	2005	196	904	2479		FALSE	FALSE	**TRUE**		196	904	___		196	904	**1474**
47	2006	177	980	2527		FALSE	FALSE	**TRUE**		177	980	___		177	980	**1513**
48	2007	980	1052	2551		FALSE	FALSE	**TRUE**		980	1052	___		980	1052	**1551**
49	2008	849	1137	1589		FALSE	FALSE	FALSE		849	1137	1589		849	1137	1589
50	2009	845	719	1552		FALSE	FALSE	FALSE		845	719	1552		845	719	1552