International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

A Deviation Revealing Approach for Data Cleaning

Darshanaben Dipakkumar Pandya[1], Dr. Sanjay Chaudhary[2], Dr. Sanjay Gaur[3]
[1]Research Scholar, Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.
[2]Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.
[3]Research Associate Professor, Department of Computer Science & Engineering, Jaipur Engineering College and Research Center, Jaipur.

Abstract: Information superiority is significant to organizations with the use of data mining; Anomalous data values detection is a most important step in many data related applications. Anomalous data make the performance of data analysis difficult. The presence of anomalous data value can also pose serious problems for researchers. In fact, in appropriate handling of the Anomalous data values in the analysis may introduce bias and can result in misleading conclusions being drawn from a research study and can also limit the generalize ability of the research findings. There are numerous techniques for Anomalous data detection, while using Inliers and Outlier techniques and their different measures in data mining. This article introduces anomalous data detection algorithm that should be used in data mining systems. Basic approaches currently used for solving this Anomalous data values finding, problem are considered, and their results are discussed using table.

Keywords: Anomalous data, Detection, Data Mining, Deviation Algorithm.

I. INTRODUCTION

In many data analysis tasks, a large number of variables are recorded or sampled. The outlier is introduced as an observation that differs too much from the other observations in the data set, which raises the suspicion that it was generated by a mechanism different from other observations. Although abnormal values are often considered to be an error or noise, they may contain important information. The outlier is a value in a data set very far from an initial model. An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the method of detection applied. However, some definitions are considered general enough to deal with various types of data and methods. The following figure shows the abnormal values in the data set.

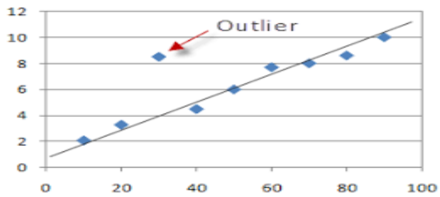

Fig.1. Diagram of outlier example in data set.

While In the database, Inliers is a data value that is within a statistical data distribution and is also considered an error. In the database, the additional data in a data set is an observation or a subset of observations, not necessarily all zeros, which appears to be irregular with the set of data remaining in the database. For example: Consider the following example as a usual occurrence of data values: 0, 0, 0, 0.02, 0.06, 0.70,1, 75, 1.15, 1, 71, 2, 44 and 3.32. Here the first three observations can be treated as continuous errors; the next three observations can be treated as early failures. The below figure shows inliers in the data set.

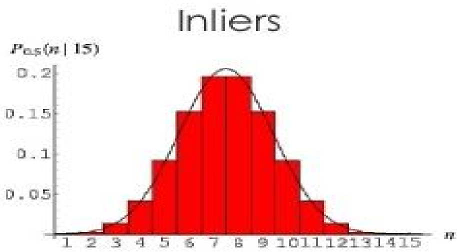

Fig.2. Diagram of inliers example in data set.

II. BACKGROUND ON ANOMALOUS DATA

In this research paper study, a technique is discussed which provides an approach to find out anomalous values from a real dataset with considerable anomalous values. Therefore, the objective of this method is to discover the outliers, inliers and notify the inliers, outlier's records completely according to particular criteria.

Arti. Rhabia, K. Muralidharan [1] are the persons who have introduced Inliers detection in Pareto distribution in various ways. K. Muralidharan [2] he introduced theory of inliers modeling and their various applications. Gaur, Sanjay and Dulawat M.S [3] are the persons who discussed about univariate Analysis for Data Preparation in context of Missing Values. Gaur, Sanjay and Dulawat, M.S [4] applied A Closest Fit Approach to Missing Attribute Values in Data Mining. Buck and S.F [5] convoluted a technique of the estimation of missing values for multivariate data suitable for use with an electronic computer. Kim, J.O., and Curry, J [6] who give the dealing of missing data in multivariate investigation", Social Methods and Research. Rubin, D.B [7] will recover Inference and missing data, Sharma, Swati and Gaur, Sanjay [8] Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining. Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori [9] are the authors who introduced Statistical outlier detection using direct density ratio estimation. C. Aggarwal and P. Yu [10] are the persons who discussed Outlier Detection for High Dimensional Data. Z. He, X. Xu and S. Deng [11] are researchers who introduced "discovering Cluster based Local Outliers".

III.PROPOSED APPROACH

As there can be reviewed several different ways of detecting anomalous data, here propose a technique which is a combination of numerical and data mining. Then apply deviation Anomaly Method approach algorithm to detect the anomalous data just like Inliers, Outliers and missing data from the dataset.

A. Deviation Anomaly Method

The proposed method is based on finding anomalous data value from the data set by the Deviation Anomaly method. In general, this method is search of anomalous data value which is very close to the true mean of the attribute. If found anomalous data in the data set then notify them as inliers, outliers depending on algorithm criteria.

1) *Algorithm*

a) *Step 1:* Select a dataset on which anomalous data detection is to be performed from the database.

b) *Step 2:* Determine the Mean ($\bar{x}$) of the data set using addition of all the data points indicated by (Σxi) and dividing it by the Number of the data points (n) using below formula.

$$\bar{x} = \frac{\Sigma x_i}{n}$$

c) *Step 3:* Calculation of Subtraction of the Mean ($\bar{x}$) from each data point and then after Square the result .

d) *Step 4:* Calculate the Mean of the result again.

e) *Step 5:* Determine the square root of that Mean to obtain the Standard Deviation (S.D) from the dataset.

f) *Step 6:* Determine Result using Standard Deviation (S.D) multiply by 1.5. Result = (S.D) * 1.5

g) *Step 7:* Calculate Lower control boundary A using the result of step 6 Subtracting from the mean of the actual data set (step 1) to get the Lower control boundary. Lower control boundary A = Mean ($\bar{x}$) – Result.

h) *Step 8:* Calculate Upper control boundary B using the result of step 6 adding from the mean of the actual data set (step 1) to get the Upper control boundary. Upper control boundary B = Mean ($\bar{x}$) + Result.

i) *Step 9:* If a data value is less than the Lower control boundary A then it is considered as inliers or greater than the Upper control boundary B, it is considered an outlier in dataset.

j) *Step 10:* If the step 9 conditions is true, then eliminate the data entry having anomalous data permanently from the dataset.

k) *Step 7:* finished.

Stop.

The below TABLE I indicates the Deviation Anomaly Method using real database. The real data set is taken from

TABLE I

A Deviation Revealing Approach for Data cleaning

Dataset Average U.S. Retail Fuel Prices, from date Apr-2000 to Jan -2015, (Dollars per Gasoline Missing)

Standard Dataset					Square Data values			Recovered Dataset		
SN	DATE	GASOLINE	DIESEL	ELECTRICITY	GASOLINE	DIESEL	ELECTRICITY	GASOLINE	DIESEL	ELECTRICITY
1	Apr-2000	1.52	1.29	0.81	1.29	1.73	0.06	1.52	1.29	INLIER
2	Oct-2000	1.54	1.46	0.84	1.23	1.32	0.05	1.54	1.46	0.84
3	Jun-2001	1.68	1.37	0.90	0.94	1.53	0.03	1.68	1.37	0.90
4	Oct-2001	1.27	1.19	0.88	1.92	2.00	0.03	INLIER	INLIER	0.88
5	Feb-2002	1.11	1.04	0.81	2.38	2.45	0.06	INLIER	INLIER	INLIER
6	Apr-2002	1.40	1.19	0.83	1.55	2.00	0.05	1.40	INLIER	INLIER
7	Jul-2002	1.41	1.18	0.87	1.54	2.02	0.03	1.41	INLIER	0.87
8	Oct-2002	1.44	1.35	0.84	1.46	1.57	0.05	1.44	1.35	0.84
9	Feb-2003	1.61	1.50	0.79	1.09	1.22	0.07	1.61	1.50	INLIER
10	Dec-2003	1.48	1.34	0.82	1.38	1.61	0.06	1.48	1.34	INLIER
11	Mar-2004	1.74	1.47	0.85	0.83	1.29	0.04	1.74	1.47	0.85
12	Jun-2004	1.99	1.55	0.92	0.44	1.12	0.02	1.99	1.55	0.92
13	Nov-2004	1.97	1.93	0.89	0.46	0.46	0.03	1.97	1.93	0.89
14	Mar-2005	2.11	2.03	0.88	0.29	0.33	0.03	2.11	2.03	0.88
15	Sep-2005	2.77	2.54	0.98	0.01	0.00	0.01	2.77	2.54	0.98
16	Jan-2006	2.23	2.32	0.95	0.18	0.08	0.01	2.23	2.32	0.95
17	May-2006	2.84	2.69	1.05	0.03	0.01	0.00	2.84	2.69	1.05
18	Sep-2006	2.22	2.37	1.08	0.19	0.06	0.00	2.22	2.37	1.08
19	Feb-2007	2.30	2.37	0.98	0.12	0.05	0.01	2.30	2.37	0.98
20	Jul-2007	3.03	2.67	1.10	0.14	0.00	0.00	3.03	2.67	1.10
21	Oct-2007	2.76	2.81	1.07	0.01	0.04	0.00	2.76	2.81	1.07
22	Jan-2008	2.99	3.05	1.01	0.12	0.20	0.00	2.99	3.05	1.01
23	Apr-2008	3.43	3.71	1.08	0.61	1.22	0.00	3.43	3.71	1.08
24	Jul-2008	3.91	4.22	1.19	1.59	2.60	0.02	OUTLIER	OUTLIER	1.19
25	Oct-2008	3.04	3.27	1.17	0.15	0.44	0.01	3.04	3.27	1.17
26	Jan-2009	1.86	2.19	1.11	0.62	0.17	0.00	1.86	2.19	1.11
27	Apr-2009	2.02	2.04	1.14	0.40	0.32	0.01	2.02	2.04	1.14
28	Jul-2009	2.44	2.27	1.19	0.04	0.11	0.02	2.44	2.27	1.19
29	Oct-2009	2.64	2.50	1.12	0.00	0.01	0.00	2.64	2.50	1.12
30	Jan-2010	2.65	2.57	1.08	0.00	0.00	0.00	2.65	2.57	1.08
31	Apr-2010	2.84	2.71	1.16	0.04	0.01	0.01	2.84	2.71	1.16
32	Jul-2010	2.71	2.65	1.19	0.00	0.00	0.02	2.71	2.65	1.19
33	Oct-2010	2.78	2.75	1.18	0.02	0.02	0.01	2.78	2.75	1.18
34	Jan-2011	3.08	3.09	1.10	0.18	0.23	0.00	3.08	3.09	1.10
35	Apr-2011	3.69	3.62	1.18	1.08	1.03	0.02	3.69	3.62	1.18
36	Jul-2011	3.68	3.54	1.20	1.06	0.87	0.02	3.68	3.54	1.20
37	Sep-2011	3.46	3.42	1.20	0.66	0.66	0.02	3.46	3.42	1.20
38	Jan-2012	3.37	3.47	1.14	0.52	0.75	0.01	3.37	3.47	1.14
39	Mar-2012	3.89	3.71	1.18	1.54	1.22	0.01	3.89	3.71	1.18
40	Jul-2012	3.52	3.36	1.19	0.76	0.57	0.02	3.52	3.36	1.19

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

41	Sep-2012	3.82	3.70	1.19		1.37	1.20	0.02		3.82	3.70	1.19
42	Jan-2013	3.29	3.55	1.14		0.41	0.89	0.01		3.29	3.55	1.14
43	Mar-2013	3.59	3.58	1.18		0.88	0.95	0.01		3.59	3.58	1.18
44	Jul-2013	3.65	3.50	1.25		1.00	0.80	0.04		3.65	3.50	1.25
45	Oct-2013	3.45	3.51	1.22		0.64	0.82	0.03		3.45	3.51	1.22
46	Jan-2014	3.34	3.49	1.15		0.48	0.78	0.01		3.34	3.49	1.15
47	Apr-2014	3.65	3.56	1.19		1.00	0.91	0.02		3.65	3.56	1.19
48	Jul-2014	3.70	3.51	1.19		1.10	0.82	0.02		3.70	3.51	1.19
49	Oct-2014	3.34	3.38	1.24		0.48	0.60	0.03		3.34	3.38	1.24
50	Jan-2015	2.30	2.75	1.27		0.12	0.02	0.04		2.30	2.75	1.27
	MEAN	2.65	2.61	1.06		0.69	0.78	0.02		2.69	2.70	1.09
	S.D	0.84	0.89	0.15		0.60	0.71	0.02		0.79	0.79	0.13

Source: www.earth_policy.com

IV. EXPERIMENTAL RESULTS

There can be experimental data which has been made by introducing some anomalous values in the real data set. The below table 1 shows deviation anomaly method of the dataset with outlier, inliers. here must erase the anomalous data entry and save both the dataset i.e. with anomalous data entry and without anomalous data entry and run further the deviation anomaly approach to do the analysis of the data and evaluate the sum of points to the value in each case.

Once evaluating both the values will compare them to check whether the existence of anomalous values increases the sum or not, if it increases then it must be removed. If a data value is less than the Lower control boundary A and greater than the Upper control boundary B, then remove the data entry having anomalous data values permanently from the dataset. Hence delete the anomalous data values permanently from the dataset, because this entry is not at all useful and distorting our original dataset.

V. CONCLUSIONS

The conclusion lies in the reality that anomalous data are usually the redundant entries which always alter the data in one or the other form and misreports the distribution of the data. Sometimes it becomes essential to keep even the anomalous data entries because they play an important role in the data but in our case achieving and our main objective is to discovering anomalous entries and i.e. to delete the anomalous data entries from database.

Proposed approach provides suitable consolidated report using data relative attributes of the database. It is observable that standards in the relative attribute or dependent attribute have certain correlations in the database. Furthermore the more work can be undertaken to identify the correlation between the attributes, which in turn shall help in discovery of anomalous values. One can also laid the emphasis on working upon the said research as a basis and evolve more types of patterns and distribution of values in the attribute for discovering anomalous values and its implications.

REFERENCES

[1] Arti. Khabia, K. Muralidharan, Inliers detection in Pareto distribution, Journal of Interdisciplinary Mathematics Vol. 15 (2012), No. 4 & 5, pp. 261–274.
[2] K. Muralidharan, theory of inliers modeling and applications, University of Bedfordshire, 2011.
[3] Dulawat and Gaur Sanjay , M.S., Univariate Analysis for the data research of Missing Values , Journal of Mathematical and Computer Sciences, Vol.-1, No. 5, pp. 628-635(2010).
[4] Gaur, Sanjay and Dulawat, M.S., A Closest Fit Approach to Missing Attribute Values in Data Mining, International Journal of advances in Science and Technology, Vol.-2, issue-4, (2011).
[5] S.F. and Buck, the technique of the estimate of missing values for the multivariate information suitable for use with an electronic computer, Series B, Vol-2, pp. 302-306(1960)
[6] Curry, J. and Kim, J.O., “The missing value treatment of the information in multivariate analysis”, Social Research and Methods, Vol. 6, pp. 215-240, 1977.
[7] Rubin, D.B., Inference and missing data, Biometrika, 63, pp. 581-592(1976).
[8] Sharma, Swati and Gaur, Sanjay, Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining”, International Journal Of Advanced Research In Computer Science, Vol. - 4(11), pp 214-217 (2013).
[9] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori (2009), Statistical outlier detection using direct density ratio estimation
[10] C. Aggarwal and P. Yu, “Outlier Detection for High Dimensional Data”. International Conference on Management of Data, Vol.-30, No.- 2, Pp-37– 46, May 2001.
[11] Z. He, X. Xu and S. Deng, “Discovering Cluster based Local Outliers”. Pattern Recognition Letters, Vol. - 24, No. - 9-10, Pp. 1641 – 1650, June 2003.