ISSN: 2321-7782 (Online)

Volume 3, Issue 5, May 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Recognisation of Outlier using Partitioning of Dataset for Large Scale Database

Bokare Madhav M[1]
Research Scholar
Priyadarshini Institute of Engineering & Technology,
Nagpur – India

Dr. Vilas M Thakare[2]
Professor
Department of CS, SGB Amravati University, Amravati.
Amravati – India

***Abstract:** A control chart is a statistical tool used to distinguish between variation in a process resulting from common causes and variation resulting from special causes. In this paper we used CC Technique to determine outliers. It presents a graphic display of process stability or instability over time. Every process has variation. Existing studies in data mining mostly focus on finding patterns in large datasets and further using it for organizational decision making. However, finding such exceptions and outliers has not yet received as much attention in the data mining field as some other topics have, such as association rules, classification and clustering.*

***Keywords:** Outliers, CCT, Centreline ,Charting, Control Limits.*

I. INTRODUCTION

In most databases analyzed, it is often found structural changes (outliers) that may be associated with unexpected events, as measurement and sample record errors. These changes are known as anomalies, distortions, aberrations and can interfere in the results of the statistical analyzes of the data.

In order to start working on data analysis, the first step is to observe the data behavior. Thus, analyzing disproportionate values, the possible outliers. These can be easily identified, but there are many exceptions due to anomalies. When expressed in a graph, for instance, outliers can occur in many different settings, from simple data collection and tabulation errors, or due to existing phenomena from which data are collected.

Figure 1 shows two samples easily defined as outliers and a possible one, the problem in cases like this is how define if the data can be classified as an outlier or not.

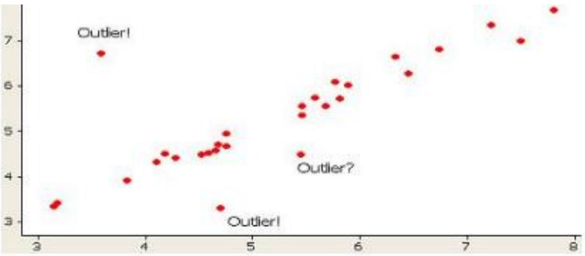


Figure 1 Outlier Detection Example

In sonic profile data, as in most cases, there is also outliers, and as stated before it is a given outlier or not we should go check the possible causes of its appearance, this case can be read error probe or malformation of the well at that point due [1]. Correctly set control limits in control charts is one of the main conditions for successful application of statistical process control and for meeting its basic goal, i.e. verifying statistical stability of the analyzed process.

Time series analysis is a part of many statistical software packages. But only some of them offer the outliers analysis (i.e. methodology for detection and assessment of possible influence of outliers).

Various accidents have occurred in underground mines worldwide in the previous years. Three accidents occurred in 1891, 1956 and 1958 in different mines within the Springhill coal field due to fire, explosion and earth quake. A total of 238 human lives were lost in the three incidents. Arguably the worst ever mine disaster in the world took place on April 26, 1942 in Benxihu Colliery, located at Benxi, Liaoning. In a very recent accident in November 2009, at least 104 miners were killed.

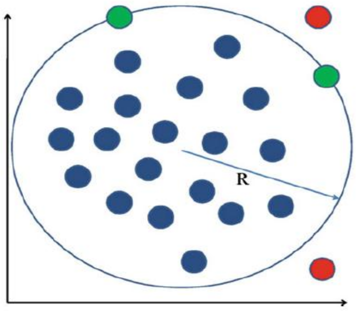


Figure 2 Outliers in 2D data. The points, outside the circle of radius

The accident was caused by a methane explosion followed by a coal dust explosion. Amethane blast at a Kemerovo coal mine killed 21miners in 2005. Since 1978, 20 mining accidents have occurred in Poland. Even with the down trend in the fatalities and accidents, 21,351 people were injured between the year 1991 and 1999. In 1972, 91 people lost their lives in Sunshine silver mine at Kellogg Idaho. In 2006, 47 out of 72 miners lost their lives in coal mining. Majority of these deaths occurred at Kentucky and West Virginia. As recent as 5th April 2010, 29 valuable lives were claimed in Upper Big Branch mine disaster at Raleigh County, Montcoal, West Virginia [4].

II. Types of Control Charts

There are two main categories of Control Charts, those that display attribute data, and those that display variables data.

Attribute Data

This category of Control Chart displays data that result from accounting the number of occurrences or items in a single category of similar items or occurrences. These “count” data may be expressed as pass/fail, yes/no, or presence/absence of a defect.

Variables Data

This category of Control Chart displays values resulting from the measurement of a continuous variable. Examples of variables data are elapsed time, temperature, and radiation dose. While these two categories encompass a number of different types of Control Chart there are three types that will work for the majority of the data analysis cases you will encounter. In this module, we will study the construction and application in these three types of Control Charts.

- X-Bar and R Chart
- Individual X and Moving Range Chart for Variables Data
- Individual X and Moving Range Chart for Attribute Data

In this module, we will study only the Individual X and Moving Range Control Chart for handling attributes data, although there are several others that could be used, such as the np, p, c, and charts. These other charts require an understanding of probability distribution theory and specific control limit calculation formulas which will not be covered here. To avoid the possibility of generating faulty results by improperly using these charts, we recommend that you stick with the Individual X and Moving Range chart for attribute data.

The following six types of charts will not be covered in this module:

- X-Bar and S Chart
- u Char
- Median X and R Chart
- p Chart
- c Chart
- np Char

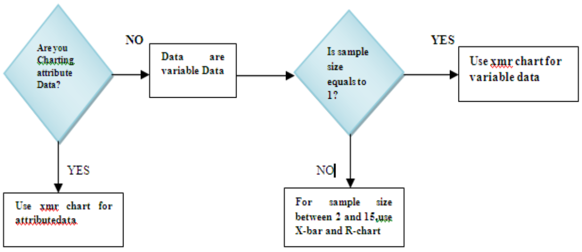

Are you Charting attribute Data?
NO
Data are variable Data
Is sample size equals to 1?
YES
Use xmr chart for variable data
YES
Use xmr chart for attributedata
NO
For sample size between 2 and 15,use X-bar and R-chart

Figure 3 Control Chart Decision Tree

The control chart decision tree is used here to charting attribute data if data id attribute data then it will use the Xmr chart for attribute data. If data is not attribute data then it will choose variable data. After this checking the sample data is equal to 1 or not. If sample data is equal to 1 then chooses the Xmr for variable data. If not then chooses the sample size between 2 to 15 for X-bar and R-chart.

III. Types of Outliers

In general, outliers can be classified into three categories, namely global outliers, contextual outliers and collective outliers [5].

1. Global outlier

In a given data set, a data object is a global outlier if it deviates significantly from the rest of the data set. Global outliers are sometimes called point anomalies and are the simplest type of outliers .Most outlier detection methods are aimed at finding global outlier .For example, Intrusion detection in computer networks [3].

2. Contextual Outliers

In a given data set, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context .Therefore in this kind of outlier; the context has to be specified as part of the problem definition [3]

3. Collective Outliers

A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers. Detection of collective outliers, consider not only behavior of individual objects, but also that of groups of objects. Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects [6].

IV. Elements of Control Chart

Each Control Chart actually consists of two graphs, an upper and a lower, which are described below under plotting areas. A Control Chart is made up of eight elements.

1. ***Title.*** The title briefly describes the information which is displayed.
2. ***Legend.*** This is information on how and when the data were collected.
3. ***Data Collection Section.*** The counts or measurements are recorded in the data collection section of the Control Chart prior to being graphed.

4. ***Plotting Areas.*** A Control Chart has two areas—an upper graph and a lower graph—where the data is plotted.

5. The upper graph plots either the individual values, in the case of an Individual X and Moving Range chart, or the average (mean value) of the sample or subgroup in the case of an X-Bar and R chart.

   a. The lower graph plots the moving range for Individual X and Moving Range charts, or the range of values found in the subgroups for X-Bar and R charts.

   b. Vertical or Y-Axis. This axis reflects the magnitude of the data collected.

The Y-axis shows the scale of the measurement for variables data, or the count (frequency) or percentage of occurrence of an event for attribute data.

6. ***Horizontal or X-Axis.*** This axis displays the chronological order in which the data were collected.

7. ***Control Limits.*** Control limits are set at a distance of 3 sigma above and 3 sigma below the centerline. They indicate variation from the centerline and are calculated by using the actual values plotted on the Control Chart graphs.

8. ***Centerline.*** This line is drawn at the average or mean value of all the plotted data. The upper and lower graphs each have a separate centerline.

The existence of these types of outliers mean that some of the observations at a sensor node are anomalous with respect to the rest of data, as shown in Fig. 4. The Local Outliers are also known as First Order Outliers.

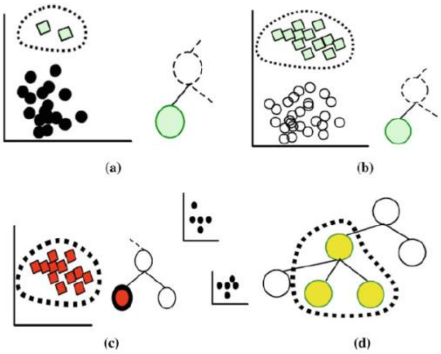


Figure 4 first order outliers. Some of the measurements are anomalous with respect to others. In the plot, *squares* represent the abnormal measurements. b First order epoch outliers (Type-4 local outliers). c Second order external outliers. All measurements of a sensor node are anomalous with respect to neighboring nodes. d Third order external outliers. A subset/sub tree of nodes is anomalous with respect to neighboring nodes in the network.

The First order outliers are further classified into following categories: Type 1 or Incidental absolute errors/outliers are isolated (one-time spike) or very short sequence of extreme high or low values. For example a temperature of 0 degrees in a desert during the day time. These outliers can be identified by using a pre-defined threshold. Type 2 or Clustered Absolute Outliers are a continuous sequence of Type 1 outliers. Type 3 or Random Errors/outliers are indicated by observations not falling within the threshold of the normal data [4].

V. Implementations and Discussions

The steps for calculating and plotting an X-Bar and R Control Chart for Variables Data:

The X-Bar (arithmetic mean) and R (range) Control Chart is used with variables data when subgroup or sample size is between 2 and 15. The steps for constructing this type of Control Chart are:

Step 1 - Determine the data to be collected. Decide what questions about the process you plan to answer. Refer to the Data Collection module for information on how this is done.

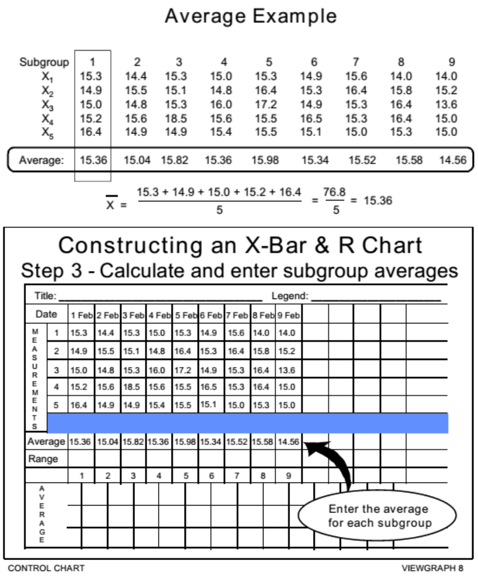
Step 2 - Collect and enter the data by subgroup. A subgroup is made up of variables data that represent a characteristic of a product produced by a process. The sample size relates to how large the subgroups are. Enter the individual subgroup measurements in time sequence in the portion of the data collection section of the Control Chart labeled MEASUREMENTS.

STEP 3 - Calculate and enter the average for each subgroup. Use the formula below to calculate the average (mean) for each subgroup and enter it on the line Labeled Average in the data collection section (Viewgraph 8)

$$\bar{x} = \frac{x1 + x2 + x3 + \cdots..xn}{n}$$

Where: $\bar{x}$ = The average of the measurements within each subgroup , Xi= The individual measurement s within a subgroup

n =The number of measurements within a subgroup



Average Example

Subgroup	1	2	3	4	5	6	7	8	9
X_1	15.3	14.4	15.3	15.0	15.3	14.9	15.6	14.0	14.0
X_2	14.9	15.5	15.1	14.8	16.4	15.3	16.4	15.8	15.2
X_3	15.0	14.8	15.3	16.0	17.2	14.9	15.3	16.4	13.6
X_4	15.2	15.6	18.5	15.6	15.5	16.5	15.3	16.4	15.0
X_5	16.4	14.9	14.9	15.4	15.5	15.1	15.0	15.3	15.0
Average:	15.36	15.04	15.82	15.36	15.98	15.34	15.52	15.58	14.56

$$\bar{X} = \frac{15.3 + 14.9 + 15.0 + 15.2 + 16.4}{5} = \frac{76.8}{5} = 15.36$$

Constructing an X-Bar & R Chart

Step 3 - Calculate and enter subgroup averages

Title: ______ Legend: ______

Date		1 Feb	2 Feb	3 Feb	4 Feb	5 Feb	6 Feb	7 Feb	8 Feb	9 Feb
MEASUREMENTS	1	15.3	14.4	15.3	15.0	15.3	14.9	15.6	14.0	14.0
	2	14.9	15.5	15.1	14.8	16.4	15.3	16.4	15.8	15.2
	3	15.0	14.8	15.3	16.0	17.2	14.9	15.3	16.4	13.6
	4	15.2	15.6	18.5	15.6	15.5	16.5	15.3	16.4	15.0
	5	16.4	14.9	14.9	15.4	15.5	15.1	15.0	15.3	15.0
Average		15.36	15.04	15.82	15.36	15.98	15.34	15.52	15.58	14.56
Range										
		1	2	3	4	5	6	7	8	9
AVERAGE										

Enter the average for each subgroup

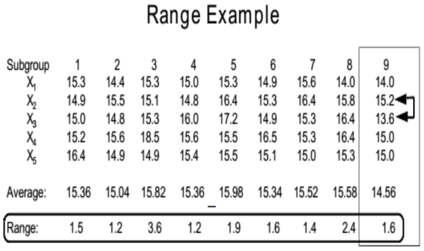
CONTROL CHART

VIEWGRAPH 8

Figure 5 Viewgraph 8

Step 4 - Calculate and enter the range for each subgroup. Use the following formula to calculate the range (R) for each subgroup. Enter the range for each subgroup on the line labeled Range in the data collection section (Viewgraph 9).

RANGE = (Largest Value in each Subgroup) - (Smallest Value in each Subgroup)



Range Example

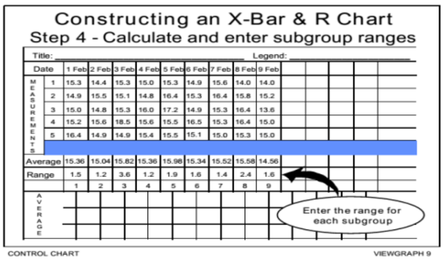
Subgroup	1	2	3	4	5	6	7	8	9
X_1	15.3	14.4	15.3	15.0	15.3	14.9	15.6	14.0	14.0
X_2	14.9	15.5	15.1	14.8	16.4	15.3	16.4	15.8	15.2
X_3	15.0	14.8	15.3	16.0	17.2	14.9	15.3	16.4	13.6
X_4	15.2	15.6	18.5	15.6	15.5	16.5	15.3	16.4	15.0
X_5	16.4	14.9	14.9	15.4	15.5	15.1	15.0	15.3	15.0
Average:	15.36	15.04	15.82	15.36	15.98	15.34	15.52	15.58	14.56
Range:	1.5	1.2	3.6	1.2	1.9	1.6	1.4	2.4	1.6

Step 5 - Calculate the grand mean of the subgroup’s average. The grand mean of the subgroup’s average (X-Bar) becomes the centerline for the upper plot.

Where: x(bar bar)= The grand mean of all the individual subgroup averages, X (bar)= The average for each subgroup

k =The number of subgroups

$$\bar{\bar{x}} = \frac{15.36 + 15.04 + 15.82 + 15.36 + 15.98 + 15.34 + 15.52 + 15.58 + 14.56}{9} = \frac{138.56}{9} = 15.40$$



Constructing an X-Bar & R Chart

Step 4 - Calculate and enter subgroup ranges

Title: ______ Legend: ______

Date		1 Feb	2 Feb	3 Feb	4 Feb	5 Feb	6 Feb	7 Feb	8 Feb	9 Feb
MEASUREMENTS	1	15.3	14.4	15.3	15.0	15.3	14.9	15.6	14.0	14.0
	2	14.9	15.5	15.1	14.8	16.4	15.3	16.4	15.8	15.2
	3	15.0	14.8	15.3	16.0	17.2	14.9	15.3	16.4	13.6
	4	15.2	15.6	18.5	15.6	15.5	16.5	15.3	16.4	15.0
	5	16.4	14.9	14.9	15.4	15.5	15.1	15.0	15.3	15.0
Average		15.36	15.04	15.82	15.36	15.98	15.34	15.52	15.58	14.56
Range		1.5	1.2	3.6	1.2	1.9	1.6	1.4	2.4	1.6
		1	2	3	4	5	6	7	8	9

Figure 6 Viewgraph 9

Step 6 - Calculate the average of the subgroup ranges. The average of all subgroups becomes the centerline for the lower plotting area.

$$\bar{R} = \frac{R1+R2+R3+\cdots RK}{K}$$

Where: Ri=The individual range for each subgroup , $\bar{R}$=The average of the ranges for all subgroups ,k =The number of subgroups.

Average of Range Example

$$\bar{R} = \frac{1.5+1.2+3.6+1.2+1.9+1.6+1.4+2.4+1.6}{9}$$

$$= \frac{16.4}{9} = 1.8$$

Step 7 - Calculate the upper control limit (UCL) and lower control limit (LCL) for the averages of the subgroups. At this point, your chart will look like a Run Chart. Now, however, the uniqueness of the Control Chart becomes evident as you calculate the control limits. Control limits define the parameters for determining whether a process is in statistical control. To find the X-Bar control limits, use the following formula.

$$UCLx = \bar{\bar{x}} + A2\bar{R}$$

$$LCLx = \bar{\bar{x}} - A2\bar{R}$$

Upper and Lower Control Limits Example

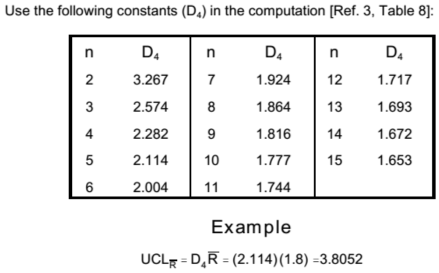
$$UCL_{\bar{x}} = \bar{\bar{x}} + A2\bar{R} = (15.40) + (0.577)(1.8) = 16.4386$$

$$UCL_{\bar{x}} = \bar{\bar{x}} + A2\bar{R} = (15.40) - (0.577)(1.8) = 14.36614$$

Step 8 - Calculate the upper control limit for the ranges. When the subgroup or sample size (n) is less than 7, there is no lower control limit. To find the upper control limit for the ranges, use the formula:

$$UCL_{\bar{R}} = D_4\, \bar{R}$$

$$LCL_{\bar{R}} = D_3\, \bar{R}\ (For\ sungroups \geq 7)$$



Use the following constants (D_4) in the computation [Ref. 3, Table 8]:

n	D_4	n	D_4	n	D_4
2	3.267	7	1.924	12	1.717
3	2.574	8	1.864	13	1.693
4	2.282	9	1.816	14	1.672
5	2.114	10	1.777	15	1.653
6	2.004	11	1.744		

Example

$UCL_{\bar{R}} = D_4\bar{R} = (2.114)(1.8) = 3.8052$

Step 9 - Select the scales and plot the control limits, centerline, and data points, in each plotting area. The scales must be determined before the data points and centerline can be plotted. Once the upper and lower control limits have been computed, the easiest way to select the scales is to have the current data take up approximately 60 percent of the vertical (Y) axis. The scales for both the upper and lower plotting areas should allow for future high or low out-of control data points. Plot each subgroup average as an individual data point in the upper plotting area. Plot individual range data points in the lower plotting area.

Step 10 - Provide the appropriate documentation. Each Control Chart should be labeled with who, what, when, where, why, and how information to describe where the data originated, when it was collected, who collected it, any identifiable equipment or work groups, sample size, and all the other things necessary for understanding and interpreting it. It is important that the legend include all of the information that clarifies what the data describe.

VI. Conclusion

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we take overview of outlier, types of outliers and outlier detection. Next, we reviews related work in outlier detection. Next, we discuss and steps of outlier detection which can be grouped into statistical-based approach, distance based approach, density-based approach, and Information theoretic-based approach. We discuss advantages and limitations of each algorithm. Finally, in implementation section, our experiments on different datasets show promising results, accurately finding outliers.

Acknowledgement

We wish to thank Dr. S. B. Thorat, Director, SSBES'S Institute of Technology and Management, Nanded. I also thankful to P. K. Butey, Head, Dept of CS, Kamala Nehru College, Nagpur. I am grateful thanks to Mr. Pawale Satish R, Lecturer at SSBES'S Institute of Technology and Management, Nanded.

References

1. Daniel Francisco Maranhão Evangelista, José Augusto Andrade Filho, Glaucio José Couri Machado, Gabriel Francisco da Silva, Suzana Leitão Russo ," OUTLIERS DETECTION USING CONTROL CHARTS FOR OIL WELLS".
2. Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, Mustafa Mat Deris, A Comparative Study for Outlier Detection Techniques in Data Mining.
3. Jiawai Han,Micheline Kamber,Jian Pei "Data Miningconcepts and techniques",morgan kaufmaan publishers,third edition.
4. Nauman Shahid, Ijaz Haider Naqvi, Saad Bin Qaisar,Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey, © Springer Science+Business Media Dordrecht 2013.
5. http://www.scribd.com/doc/17295273/Quality-Management#scribd.
6. A.SARAVANAN, DR P. NAGARAJAN, IMPLEMENTATION OF QUALITY CONTROL CHARTS IN BOTTLE MANUFACTURING INDUSTRY, A.Saravanan et al. / International Journal of Engineering Science and Technology (IJEST).
7. www.au.af.mil/au/awc/awcgate/navy/bpi_manual/mod10-control.pdf

Author(s) Profile

Mr. Madhav Bokare, M.E. (CSE), 25+ National /International/Journals publication. The area of research is Computer Networks, Data Mining. Currently working as a Head of Dept of CS at SSBES'S ITM college, Nanded.

Dr. V. M. Thakare, Ph.D. (Computer Science), M.E. (Advance Electronics), M.Sc. (Applied Electronics), Diploma in Computer Management. Member of: Institute of Engineers (Life Member), Indian Society of Technical Education ISTE (Life Member), Computer Society of India CSI (Member).