International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

A Comparative Study for Anomaly Detection in Data Mining

Fenil Shingala[1], Shalin Barot[2], Bhavesh Oza[3]

[1,2] Student, LDCE Ahmedabad
[3]Professor, Dept. of Computer Engineering, LD College of Engineering, Gujarat, India

---

Abstract - *In this paper, we will discuss some of the research we have found till and what we have concluded from that survey. We try to compare and combine three of the methods we have explored. We will work on Outlier / Anomaly Detection. Data mining is the process of extraction of data that would be of any kind and Outlier / Anomaly is detection of irrelevant data.*

Key Words: **moving averages, mean absolute deviation, DBSCAN, regression, anomaly detection, outlier detection, prediction analysis, data mining**

1.INTRODUCTION

Anomaly detection was originally proposed for intrusion detection systems to prevent cases such as burglary. Anomaly detection for IDS is normally accomplished with statistical analysis by defining boundaries, yet ought to similarly be conceivable with soft computing, and inductive learning. [1]

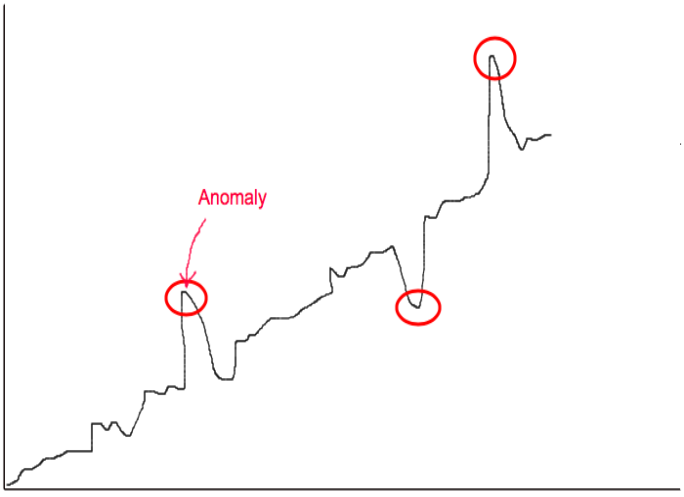


Fig -1: Anomaly detection

In data mining, anomaly detection is the identification of items, events or observations that do not comply with an expected pattern or other items in a dataset. Normally the anomalous items will mean some sort of issue, for example bank extortion, a structural defect, medical problems or mistakes in content. For example, as figure shows, in given time series there are some extreme points, which are largely deviating from other points.

Anomaly detection:

Observing datasets for certain time period and afterward analyzing patterns to find abnormal behaviors of upcoming data points detect anomalies. Here we are comparing few methods to do so. They include basic ideas of statistical mathematics and clustering techniques.

Moving averages:

In statistics, a moving average is a process of creating a series of averages of a certain portion of data sets of the entire data. These averages are called moving mean or rolling mean. [2]

But how can we do this? The answer is sliding window. First of all, we will take a series of numbers. Then decide a fixed subset size that will become a window. Start from initial index to window size and take an average of numbers. Then slide a window by removing first number and adding next number in series after current window size, and take an average. This will produce a series of averages which is called moving averages.

It can also be used for anomaly detection. After finding first moving average we can define threshold limits plus and minus to current moving average to list out anomalies.

Example:

Let's consider we are analyzing data taken from a tractor company. We take sell of each month and plot a graph. There will be three parts to this graph. Middle part; duration is number of tractors sold were of not too much or too less proportion. Other two parts will be of too much and too less sell in which so ever duration.

Let's say we're interested in finding periods during which sell is on peak, so that we can keep more number of tractors during those duration to save ourselves from shortage. Also, we need to make sure that we do not order more quantities than needed when there is less sell. This situation can be solved by observing history; that is to say look for past sell records and find out which months are on peak of sell and which months have comparatively less sell. That is exactly what moving averages do in terms of time series analysis. Figure shows a graph for tractor sales and by applying algorithm we can find out anomalies that are deviating from

given period average. So whenever graph goes to extremes it is considered as anomalous. Here, in graph such extreme points are depicted in red. It shows upper and lower extreme values of sales during a particular time, which is exactly what we seek to find out.

Moving averages can be used for measuring the trend of any series. This method is applicable to linear as well as non-linear trends. [3]

On the other hand, the trend obtained by moving averages generally is neither a straight line nor a standard curve. For this reason the trend cannot be extended for forecasting future values.

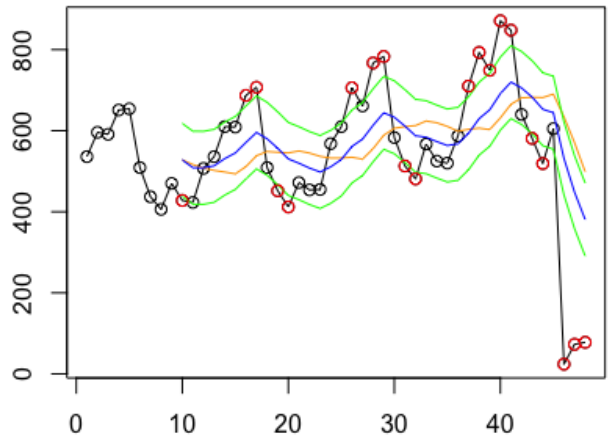


Fig -2: Moving Averages

Mean absolute deviation:

In statistics, the median absolute deviation (MAD) is a method of finding the variability of a sample that includes one variable quantity. It can also refer to the population parameter that is estimated by the MAD calculated from a sample. [4]

For a univariate data set [a1, an] the MAD is characterized as the median of the absolute deviations from the data's median:

$MAD=median|\ a_i - median\ |$

So in a nutshell, MAD is the median of absolute deviations from the data's median.

How to find MAD:

1. Find the mean/median of data
2. Find the absolute differences between each data value to the mean/median
3. Find the mean/median of these differences

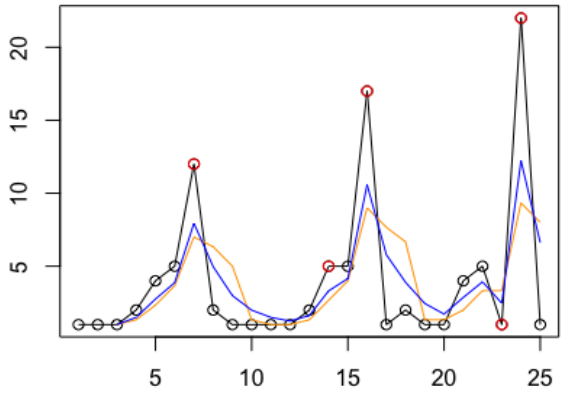
Here, note that mean or median both can be used to find MAD depending on the type of data.

MAD can be used for anomaly detection by defining thresholds for modulus of division of each point by MAD (point/MAD). This values after division gives distribution of data points with respect to their mean. Then we can easily define a threshold according to the distribution we get to extract out outliers.

Data outside the threshold are considered anomalous. It is more suitable to the types of data whose most of the points range between particular limits, and anomalies are deviating in such a way that differences from anomalous data to MAD are larger than most of the points.

Example:

Suppose a company is hosting a cloud service for which they maintain servers and serves subscribed customers with data storage. Let's say, in a hypothetical situation, company's all customer uses most of their cloud storage and also they all access it at the same time, then in this type of situation there will be too much burden on CPUs and storage devices. This type of situations are common in shopping sites mostly, where for particular given time period there is a lot of customers make transactions, such as during festivals. So in this type of situation, whenever critical situation arises, that is, when load is extreme, they can take precautionary steps in future to avoid such crises.



While this method shows promisingly advantageous mathematical side of its implementation, the same side also is also a cause of its minus point.

Merits:

- Mean deviation is broadly utilized as a part of different fields, for example, Economics, Business, Commerce or some other field of such sort.
- When comparison is required this is perhaps the best measure between at least two arrangements.
- This calculation has its base upon measurement than an estimate.
- As it is based on the deviations about an average, it gives us better measure for correlation.

Demerits:

- If Range increments on the off chance that the sample increases, average deviation additionally increments yet not in a similar proportion.
- For Sociological studies, it is practically not used. [5]

DBSCAN:

Density Based Spatial Clustering Algorithm is very useful for finding shapes based on density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based clustering algorithm. Other such an algorithm is K-means. These types of algorithms use measure of reachability and eligibility of forming a cluster. [6]

Reachability is the condition defined by parameter ε (epsilon). It stands for the maximum allowed distance between two points under consideration to be eligible for being neighbor of each other.

Eligibility to form a cluster is defined by a number. That number indicates minimum number of points required to form a cluster. [6]

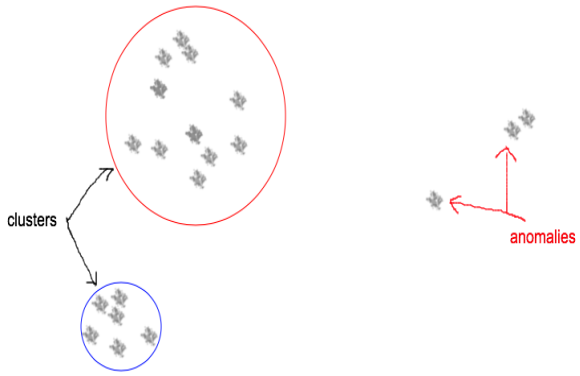
Epsilon is a parameter to be given manually for maximum distance to be considered eligible for including data point into a cluster. [7]

This clustering method creates a sort of chaining by which two distant points that are at more than ε distance will eventually be considered as neighbors. Let's understand this by an example. Suppose a random point 'a' is within ε distance to another point 'b', hence are neighbors. This point 'b' in turn is neighbor of point 'c'. By chaining, point 'a' and 'c' will also be neighbors even if they're at more than ε distance apart.

Algorithmic steps for DBSCAN clustering

Let's understand this by taking sample data points of integer values. Now, DBSCAN takes two parameters: ε (epsilon) and The minimum number of points required to form a cluster (p).

1) Start with a random starting point that has not been visited.
2) Extract the neighbor of this point using ε (All points which are within the ε distance are neighbors) by applying condition difference of each point and point currently in observation < ε.
3) If there are enough neighbors around this point then clustering process starts and point is marked as visited else this point is labeled as noise (But later on this point can become the part of the cluster).
4) If a point is found to be a part of the cluster then its ε neighbors are also the part of the cluster and the step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster are determined.
5) After the formation of the cluster, a next random unvisited point is taken to repeat the same process of clustering.
6) This process continues until all points are marked as visited.



Example:

Let's say, we are developing an algorithm for auto recognizing lone islands on the map of world or any map whatsoever. DBSCAN provides perfect solution for this type of situation. We can decide epsilon by providing minimum area for a land to fall into a category of "not a lone island". We can have these lone islands by looking at resulting outliers after applying algorithm.

Major features:

Discover clusters of arbitrary shape
Handle noise
One scan
Need density parameters[9]

Comparison:

- Unfortunately, moving averages don't work for all scenarios, especially for those in very volatile observations or those that are heavily influenced by current events. [10]
- While Mean absolute deviations has limitations that it is only applicable when there is no measurement error and when distributions are normal. [11]
- DBSCAN does not work too well when we are dealing with clusters of varying densities or with high dimensional data. [7]

Table -1: Comparison of techniques

Algorithm	Execution Time	Efficiency	Number of false alerts
Moving Averages	High	High	Less
MAD	Less	Medium	More
DBSCAN	Medium	Medium	Average

3. CONCLUSIONS

This brief paper discusses about what Anomaly detection is, and comparison of different techniques in data mining. Moving averages, Mean Absolute Deviation (MAD) and DBSCAN are most frequently used algorithms in practice. After comparison it can be said that all three of them have their own advantages and disadvantages and they can best be applied in different circumstances.

REFERENCES

[1] Hodge, V. J.; Austin, J. (2004). "A Survey of Outlier Detection Methodologies".

[2] Statistical Analysis, Ya-lun Chou, Holt International, 1975.

[3] http://www.emathzone.com/tutorials/basic-statistics/merits-and-demerits-of-moving-average-method.html

[4] Geary, R. C. (1935). The ratio of the mean deviation to the standard deviation as a test of normality. Biometrika, 27(3/4), 310–332.

[5] http://www.publishyourarticles.net/knowledge-hub/statistics/merits-and-demerits-of-mean-deviation/1096/

[6] Arlia, Domenica; Coppola, Massimo. "Experiments in Parallel Clustering with DBSCAN"

[7] http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter004/section004/blue/page003.html

[8] http://www.ques10.com/p/9286/dbscan-clustering-algorithm-with-an-example/

[9] http://www.investopedia.com/articles/trading/11/pitfalls-moving-averages.asp

[10] http://www.investopedia.com/articles/trading/11/pitfalls-moving-averages.asp

[11] http://influentialpoints.com/Training/absolute_deviations.htm