Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

Comparision of Outlier Detection Techniques using KNIME Analytics Platform

Ayush Jindal
VIT University Vellore

Nikhil Panda
VIT University Vellore

Lavanya K.
VIT University Vellore

***Abstract*— Data mining usually refers to the analysis of large data sets in order to generate new information. Outlier detection is one of the important aspects of data mining and is still gaining more relevance with time. Outlier detection has become a major area of interest for data analysts. This paper tries to provide a detailed and exclusive overview of the four outlier detection techniques that will be covering in our area of research. We have focused on the approach adopted by each technique in order to identify the outliers present in the dataset. We have taken into consideration of the key assumptions, which are used by the techniques to filter out points which are outliers When applying a given technique to a dataset, these assumptions are taken into consideration for calculating the effectiveness of the technique in a given domain. We will first discuss the basic outlier technique and then cover the variants of this technique further on.**

Keywords*— *Outlier Detection, Numeric Outlier, Z-Score, DBSCAN, Isolation Forest

I. INTRODUCTION

An outlier is defined as a data point which is anomalous from the rest of the data based on some measure. Such a point often contains useful information about the loopholes of the system described by the data and hence is a major field of research for the data analyst. The outlier detection technique finds applications in credit card fraud detection, construction domain etc. These problems typically arise in the context of very high dimensional data sets where outliers are difficult to trace. Much of the recent work focuses on finding outliers use methods which make implicit assumptions for relatively low dimensionality of the data. These methods have lower efficiency rate when the dimensionality is high, and the data becomes sparse having a steep drop in efficiency. This template provides an easier and simple understanding of the Classification based Outlier techniques. Further we identify the advantages and disadvantages that these techniques have for different domains

II. METHODOLOGY

The detection of outliers has become important in every day to day data mining applications, there are several factors that determine how to formulate the problem. There various factors which affect the problem of outlier detection can be listed as follows:

1.Nature of Input Data: The most important attribute is to know about the input data. Normally referred to as the collection of instances, the type can affect the results of the outlier detection to a great extent. In this category we are getting to handle two types of data, one that consists of one variable while if it has multiple data variables then it is called multivariable. Other types include continuous data, categorical data, spatial data and hence depending upon the data variable we can accordingly apply the required detection method.

2.Types of Outlier: An important aspect of the entire process is to have a detailed study about the type of outliers that lies in the dataset. Though Virtually impossible for large datasets one can roughly guess the type of outliers that is present in the dataset and categorize them accordingly. Depending on their behavior we have the following categories:

A (Point Outliers), B (Contextual Outliers), C (Collective Outliers), D (Real Outliers), E (Erroneous outliers).

Now let us take a deep look into each kind of the type of outliers.

•*POINT OUTLIER*: It is one of the easiest outliers to detect. This type of outliers is completely anomalous from the other data points is termed as Point Outlier. This type is the major focus of research in the outlier data field. Considering the example in which fraud detection of credit card occurs, the total amount withdrawn is greater than the upper limit for an individual's account can be termed as a Point outlier.

•*CONTEXTUAL OUTLIER*: The outlier which is completely anomalous to other data points in a dataset and is kind of similar to another dataset is termed as contextual dataset. It also determines the positioning of a certain element in the contextual basis then one can identify it as a TYPE B outlier. Applying the contextual outlier detection makes sense as these data can totally change the meaningfulness of the application domain.

•*COLLECTIVE OUTLIER*: If an individual data point is not considered anomalous but a collection of data points is suspected to be an omalous with respect to the entire dataset then it can be considered as a collective outlier. This type of outlier data is very difficult to find since the data may be distributed all over the dataset and is extremely difficult to group these and can be time consuming and cumbersome.

•*REAL OUTLIERS*: These types of outliers are of major interest of the system analyst since they normally provide something interesting that the analyst is looking for in order to discover the loopholes of the system. These observations do

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

have something interesting that helps to find the analyst something new and innovative and if they are removed somehow, we are completely left with the normal data points but this doesn't mean they are the correct values. They can't be regarded as noise rather they are the real data points. These types of outliers are the data points which have a changed base value due to the interference of noise and though these may appear as real data points but these aren't since these don't represent the corresponding dataset.

•*ERRORNEUS OUTLIERS*: These types of outliers are generally caused due to human error where the data points during the time of recording only are wrong. These types of data are also called illustrious data or erroneous data since the system plays no role in the cause of the outliers. This type of outliers is rare these days due to high grade inspection and error detection methods which can easily find these type of data points and filter them out.

There are various approaches to data outlier detection method and labels that indicate the required effort to obtain outlier free dataset:

•*SUPERVISED OUTLIER DETECTION*: Techniques that come under the category of Supervised category are assumed to be datasets that are easily available and requires a though process for the detection of outliers. The first and foremost step is to recognize the data belongs to which class. Once the class is determined one can easily filter out the dataset through a thorough data check-up even if the outlier is distributed collective outlier. But there are two major issues that may arise in supervised outlier detection; first, the anomalous instances are few so going through a process that is thorough as well as time consuming may increase the project costs as compared to the normal cost estimated. Moreover, there are many techniques that inject artificial outliers in a normal dataset to obtain a labelled training data set.

•*SEMI SUPERVISED OUTLIER DETECTION*: Techniques which are operating in the semi supervised mode, assume that the training data has certain built in functions, but are meant only for the normal class. Since they don't require labels for the outlier class, as more or less they follow the minimization technique to detect an outlier depending upon the type of outliers and are superior to the supervised approach that follows the same procedure for all the outliers irrespective of their type thus adding up expenses. For example, in space craft fault detection[17], an outlier scenario will definitely be an interior one which is certainly difficult to detect and hence requires a thorough analysis of data points to figure out the required defects or fault in the machinery whereas a simple task such as fraud detection or credit card forgery where the normal outliers can easily be detected are the ones which requires much less processing than the previous example.

•*UNSUPERVISED TECHNIQUES*: This approach makes a basic assumption that the frequency of the normal data points is much higher than the outliers. If this is not the case then this approach has a high false alarm rate. The approach also fails to detect erroneous outliers where the outliers behave like a normal data point and thus producing misguided results. Though this is the most widely used approach, this also isn't every time reliable as the two above techniques.

III. IMPLEMENTATION:

The dataset we used to test and compare the proposed outlier detection techniques is Significant Earthquakes, 1965-2016. This dataset includes a record of the date, time, location, depth, magnitude, and source of every earthquake with a reported magnitude 5.5 or higher since 1965.Some of these points in these columns have anomalies that make the through data analysis completely difficult and thus reduce system reliability. Hence, we applied four outlier detection methods to filter out outliers and ensure system reliability. Such techniques are especially useful for fraud detection where malicious attempts often differ from most nominal cases.
In order to demonstrate these techniques, we focused on finding out the outliers in terms of depth of these earthquakes.

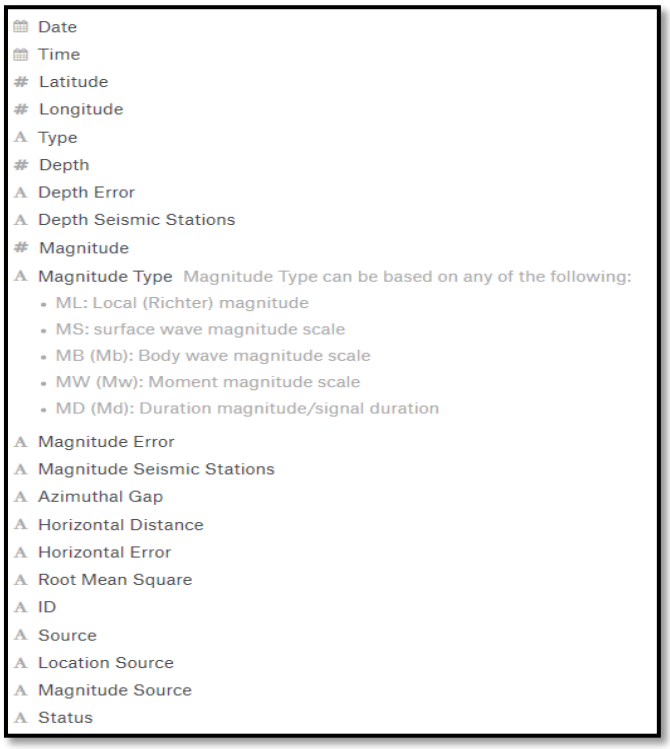


Figure 1:Attributes of Dataset used

Random Sample of Dataset

Date	Time	Latitude	Longitude	Type	Depth	Magnitude	Magnitude Type	ID	Source	Location Source	Magnitude Source	Status
12-08-1965	13:12:56	-18.108	-178.062	Earthquake	645	6.2	MW	ISCGEM851614	ISCGEM	ISCGEM	ISCGEM	Automatic
03-08-1966	01:13:43	-13.696	166.555	Earthquake	35	6.2	MW	ISCGEM849028	ISCGEM	ISCGEM	ISCGEM	Automatic
05/23/1966	08:39:45	30.081	139.992	Earthquake	20	5.7	MW	ISCGEM847725	ISCGEM	ISCGEM	ISCGEM	Automatic
01/24/1967	09:29:11	-0.922	-21.081	Earthquake	10	6.5	MW	ISCGEM840094	ISCGEM	ISCGEM	ISCGEM	Automatic
12/24/1967	21:32:33	17.543	-61.31	Earthquake	15	6.2	MW	ISCGEM828721	ISCGEM	ISCGEM	ISCGEM	Automatic
10-08-1968	14:53:38	-23.34	-66.582	Earthquake	209.3	5.6	MW	ISCGEM815732	ISCGEM	ISCGEM	ISCGEM	Automatic
11/17/1968	07:41:18	-1.238	-13.862	Earthquake	18.6	6.2	MW	ISCGEM815166	ISCGEM	ISCGEM	ISCGEM	Automatic
12/18/1968	20:03:46	-19.802	-177.609	Earthquake	376.6	5.8	MW	ISCGEM814276	ISCGEM	ISCGEM	ISCGEM	Automatic
05/30/1969	16:22:47	-32.16	-177.984	Earthquake	20	6.1	MW	ISCGEM809406	ISCGEM	ISCGEM	ISCGEM	Automatic
07/31/1969	11:23:04	52.86	-170.052	Earthquake	50	5.7	MW	ISCGEM807593	ISCGEM	ISCGEM	ISCGEM	Automatic
08/15/1969	04:32:01	43.305	147.885	Earthquake	19.2	6.3	MW	ISCGEM805834	ISCGEM	ISCGEM	ISCGEM	Automatic
10/24/1969	00:46:14	52.324	-168.696	Earthquake	27.1	5.6	MW	ISCGEM803704	ISCGEM	ISCGEM	ISCGEM	Automatic
12/25/1969	22:31:06	16.062	-59.854	Earthquake	20	6.1	MW	ISCGEM801439	ISCGEM	ISCGEM	ISCGEM	Automatic
01-10-1970	14:16:28	6.71	126.765	Earthquake	40	5.7	MW	ISCGEM799953	ISCGEM	ISCGEM	ISCGEM	Automatic
12-10-1970	04:34:41	-4.026	-80.542	Earthquake	25	7.2	MW	ISCGEM789028	ISCGEM	ISCGEM	ISCGEM	Automatic
02-06-1971	01:30:32	-49.935	-115.51	Earthquake	15	5.9	MW	ISCGEM786823	ISCGEM	ISCGEM	ISCGEM	Automatic
03/25/1971	16:19:52	38.467	142.257	Earthquake	45.9	5.7	MW	ISCGEM786621	ISCGEM	ISCGEM	ISCGEM	Automatic
09/27/1971	14:56:13	-5.822	149.452	Earthquake	110	5.7	MW	ISCGEM780659	ISCGEM	ISCGEM	ISCGEM	Automatic

Figure 2: Random Sample of Data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

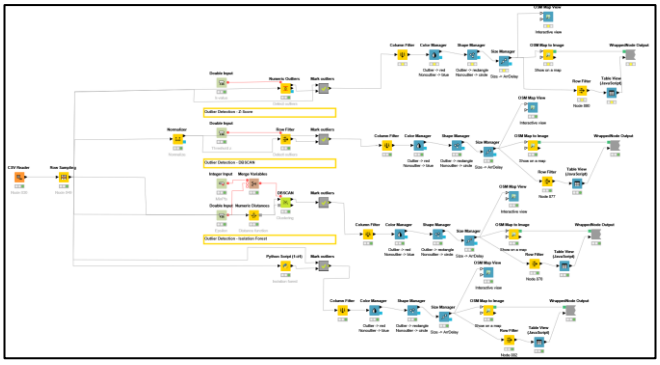


Figure 3:Workflow in KNIME

FOUR OUTLIER DETECTION TECHNIQUES

1.NUMERIC OUTLIER

This is the simplest nonparametric outlier detection method in a one-dimensional space. Here outliers are calculated by means of the IQR (Interquartile Range). The first and the third quartile (Q1, Q3) are calculated with respect to the second quartile which is therefore represented as Q2 otherwise. An outlier is thus a data point xi that lies outside the interquartile range. That is: xi >Q3 +K(IQR) or xi <Q1 - k(IQR) which in our case will be k =1.5. This technique can be easily implemented in KNIME Analytics Platform using the Numeric Outliers node which is one of the new features in KNIME Analytics Platform 5.0.

Figure 4:Map after Numeric Outlier Detection

Outlier Points

	Date	Latitude	Longitude	Depth	Magnitude
	12-09-1965	-18.108	-178.062	645	6.2
	10-08-1968	-23.34	-66.582	206.3	5.6
	12/19/1968	-19.682	-177.609	376.6	5.8
	05/13/1979	-4.054	123.149	515	5.7
	08/17/1979	-4.372	127.87	266	5.5
	09/13/1980	-4.109	127.499	239	5.6
	08/15/1983	-17.182	-174.513	181.5	5.5
	02/16/1984	36.431	70.926	207.6	6.4
	02/27/1984	-16.085	167.951	201.3	5.7
	11-06-1984	-18.666	-175.421	229.5	5.6
	07/27/1988	18.906	145.586	217.1	5.8
	05-08-1989	-23.427	-179.953	548.2	6.3
	06-11-1991	-18.289	-178.409	627.8	5.5
	02-07-1996	3.722	122.431	607.2	5.7
	12/19/1997	13.84	-68.739	162.1	6.1
	04/27/1998	-6.082	113.897	590.7	5.7
	10-11-1998	-21.04	-179.11	623.9	5.9
	09/19/1999	-5.414	146.979	238.5	5.7
	01-10-2004	-5.314	128.041	431.5	5.5
	01-01-2005	31.922	139.057	296.4	5.5
	07/10/2006	-19.937	-178.401	596.5	5.7
	09/22/2006	-20.868	-63.140	566.3	6
	02/17/2007	-4.095	154.45	403.1	5.5
	01/28/2010	-22.357	-66.712	206.4	5.9
	05-11-2013	-17.565	-175.107	212.9	6.4

Showing 1 to 25 of 25 entries

Figure 5:List of Outlier detected using Numeric Outlier Method

2.Z-SCORE

The Z-score is a parametric method to calculate the standard deviation of the data points from the sample mean assuming that the sample has a gaussian distribution. Some Python libraries like SciPy and Sci-kit Learn have easy to use functions and classes for easy implementation along with Pandas and NumPy. After applying certain transformation to the dataset, the Z-score can be calculated using the following formula:

$Z= (x-\mu)/\sigma$

Z-score is a powerful and simple method to remove outliers in low dimensional feature space but is not that effective in multi-dimensional feature space. The method's efficiency significantly drops when the distribution isn't parametric, nor does it work well with large dataset.

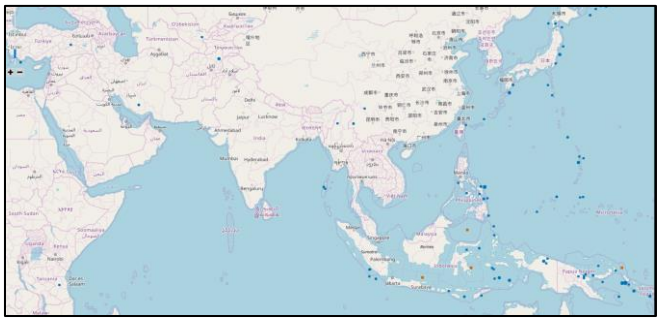


Figure 6:Map after Z-score Outlier Detection

Outlier Points

	Date	Latitude	Longitude	Depth	Magnitude
	12-09-1965	-18.108	-178.062	4.171741151897033	6.2
	05/13/1979	-4.054	123.149	3.949060229000442	5.7
	05-08-1989	-23.427	-179.953	3.453224040068403	6.3
	06-11-1991	-18.209	-178.409	4.044070756102988	5.5
	02-07-1996	3.722	122.431	3.891163186047328	5.7
	04/27/1998	-6.082	113.097	3.766868681454202	5.7
	10-11-1998	-21.04	-179.11	4.015122236126	5.9
	07/16/2006	-19.937	-178.401	3.811740326547543	5.7
	09/22/2006	-20.868	-63.149	3.825101181913385	6
	02/17/2007	-4.095	154.45	2.970006437984252	5.5

Showing 1 to 10 of 10 entries

Figure 7:List of Outlier detected using Z-Score Method

3.DBSCAN

It refers to Density Based Spatial Clustering of Applications with noise helps us to visualize and understand data better. Relationship between data in one dimensional as well as multi-dimensional feature space can be understood, and outliers can be easily detected using DBSCAN method. Even if the distribution is non-parametric it's efficiency always remains the same hence DBSCAN remains one of the most widely used methods for outlier detection. DBSCAN is a cluster-based technique where it is focused on finding the outliers on the basis of density on an n- dimensional feature space. The greater the distance of the neighboring points from the core points the greater the probability of it being a outlier.

DBSCAN method categorizes the data points into three categories:

•*Core point*: A is a core point if the neighborhood (defined by ε) contains at least the same number or more points than the parameter MinPts.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

•*Border point:* C is a border point that lies in a cluster and its neighborhood does not contain more points than MinPts, but it is still 'density reachable' by other points in the cluster.
•*Outlier:* N is an outlier point that lies in no cluster and it is not 'density reachable' nor 'density connected' to any other point. Thus, this point will have its own cluster.

There are various parameters to be considered in the DBSCAN process. The parameter value selection carries great importance in the overall efficiency of the outlier detection method. Hence the value must always be chosen taking into consideration all the points. If the value is too small, then there might be a possibility that many points might remain unclustered. But if the value is too high then the cluster might merge majority of the points into a single cluster. Sometimes it is also important to keep the size of the dataset in mind since larger the dataset the larger the value of the MinPts should be otherwise the implementation may take too much time but in general small eps values are always recommended. DBSCAN has significant role in biology, medicine, social sciences, archaeology, marketing due to its clustering approach to find outliers. The technique also carries significance in e-commerce. We can know which products the customers are looking for and can recommend the relevant products thus increasing the consumer's convenience. We can apply DBSCAN to find the cluster of products that the customer has brought. We can also use it to recommend relevant products to other customers with good reviews.

Reachability is a non-symmetric relation since, by definition, no point may be reachable from a non-core point, regardless of distance (so a non-core point may be reachable, but nothing can be reached from it!). Therefore, a further notion of connectedness is needed to formally define the extent of the clusters found by this algorithm.
Two points p and q are density-connected if there is a point of such that both p and q are density-reachable from o. Density-connectedness is symmetric.
A cluster satisfies two properties:
•All points within the cluster are mutually density-connected.
•If a point is density-reachable from any point of the cluster, it is part of the cluster as well.
Sci-kit Learn has an implementation of DBSCAN that can be used along pandas to build an outlier detection model.
Again, the first step is scaling the data, since the radius ε will define the neighborhoods along with MinPts. (Tip: a good scaler for the problem at hand can be Sci-kit Learn's Robust Scaler).
After scaling the feature space, is time to choose the spatial metric on which DBSCAN will perform the clustering. The metric must be chosen depending on the problem, Euclidean metric works well for 2 or 3 dimensions, the Manhattan metric can also be useful when dealing with higher dimensional feature spaces 4 or more dimensions.
Then, the parameter eps (ε) must be chosen accordingly to perform clustering. If ε is too big many points will be density connected, if its too small the clustering will result in many meaningless clusters. A good approach is to try values ranging from 0.25 to 0.75.
DBSCAN is also sensitive to the MinPts parameter, tuning it will completely depend on the problem at hand.

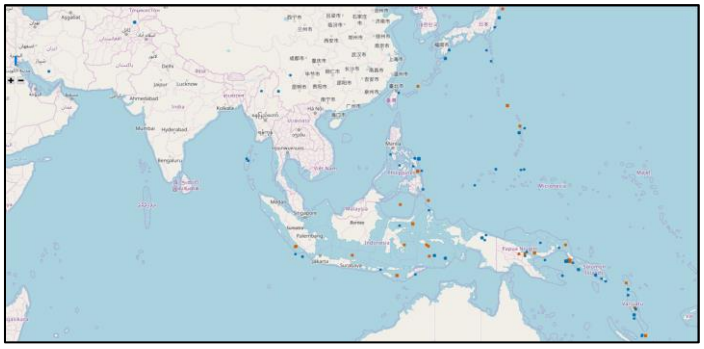


Figure 8:Map after DBSCAN Outlier Detection

Figure 9:List of Outlier detected using DBSCAN Method

4. ISOLATION FOREST:

This is a non-parametric method for large datasets for one or multi- dimensional feature space.
An important and interesting approach that this method is particularly faster is that it isolates a data point.
•A point "a" to isolate is selected randomly among the many other points.
• Another point "b" is carefully selected such that it value lies between the range and must be different from that of "a".
•If the value of "b" is greater than "a" the value of "b" becomes the new upper limit.
•If the value of "b" is less than "a" the value of "a" becomes the new upper limit.

This method is continued until we isolate an outlier. The method is purely since it is easier and more efficient to isolate an outlier than isolating a non-outlier point. Moreover, these points have lower isolation number compared to the non-outliers and hence will be faster to isolate them. The method is applied using python SKlearn library using KNIME python integration.
Finally, isolation forests are an effective method for detecting outliers in the dataset. In terms of precision, robustness, accuracy and precision isolation forest as suggested by many data analyst is the best outlier method. The memory usage and being computationally less expensive than the earlier discussed methods isolation forest not only detects the outliers but also does it faster than DBSCAN, more accurately than Z-score and with less memory usage than numeric outlier. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

method also has a high efficiency when the dimensionality factor comes into the equation. Though the method is considered supreme by many but isolation forest still suffers many disadvantages. The complexity faced during installation and debugging makes it very less adaptive to changes. That is why, inspite of having so many advantages over the others people still are not quite willing to use it.

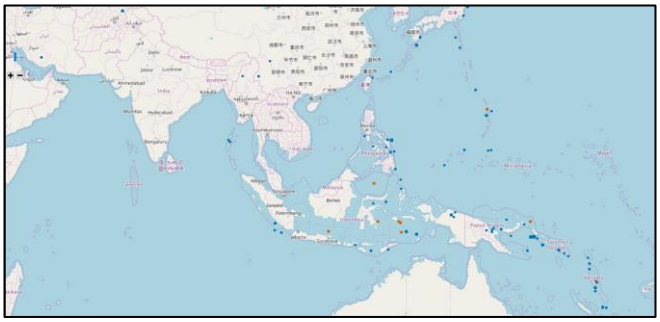
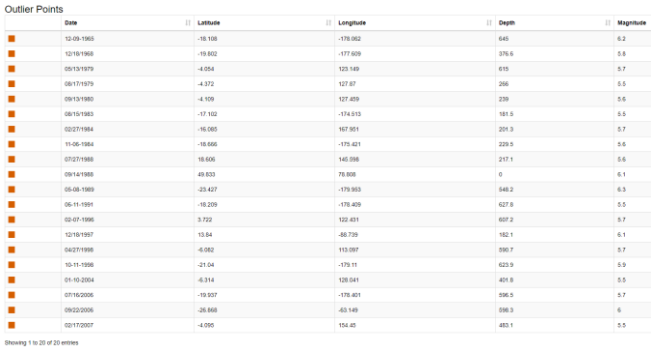


Figure 10:Map after Isolation Forest Outlier Detection



Outlier Points

	Date	Latitude	Longitude	Depth	Magnitude
■	12-09-1965	-18.108	-178.062	645	6.2
■	12/18/1968	-19.602	-177.609	376.6	5.8
■	05/13/1979	-4.054	123.149	615	5.7
■	08/17/1979	-4.372	127.87	266	5.5
■	09/13/1980	-4.109	127.459	239	5.6
■	08/15/1983	-17.102	-174.513	181.5	5.5
■	02/27/1984	-16.085	167.951	201.3	5.7
■	11-06-1984	-18.666	-175.421	229.5	5.6
■	07/27/1988	18.606	145.596	217.1	5.6
■	09/14/1988	49.833	78.806	0	6.1
■	05-08-1989	-23.427	-179.953	548.2	6.3
■	06-11-1991	-18.209	-178.409	627.8	5.5
■	02-07-1996	3.722	122.451	607.2	5.7
■	12/18/1997	13.84	-88.739	182.1	6.1
■	04/27/1998	-6.082	113.097	590.7	5.7
■	10-11-1998	-21.04	-179.11	623.9	5.9
■	01-10-2004	-6.314	128.041	401.8	5.5
■	07/16/2006	-19.937	-178.401	596.5	5.7
■	09/22/2006	-26.868	-63.149	598.3	6
■	02/17/2007	-4.095	154.45	483.1	5.5

Showing 1 to 20 of 20 entries

Figure 11:List of Outlier detected using Isolation Forest Method

IV. CONCLUSION:

Outlier detection technique	Normality Assume	Dimension	Big data	Required Pre-processing	Parameter
Numeric Outlier	NO	1	NO		IQR multiplier k
Z-score	YES	1-> low	NO	Normalize	Threshold z
DBSCAN	NO	1>multi	NO	Normalize, Calculate dist. matrix	Required number of neighbors MinPts, distance €, distance measure.
Isolation forest	NO	1>multi	YES		Estimated % of outliers

Figure 12:Comparasion on above Outlier Detection Techniques

A comparison between above four implemented outlier detection technique has been done in Figure 12 on various factors like Dimension of Data, can be used with Big Data or not, Processing required, Normal Distribution assumption is taken or not and Parameters that are needed to be tuned for working of the algorithm so that the reader has a good understanding of the pros and cons of each outlier detection and when to use it.

The process of outlier detection has applications in numerous domains, where it is desirable to determine interesting and unusual events in the activity which generates such data. The core of all outlier detection methods is the creation of a outlier free, statistical or algorithmic model which characterizes the normal behavior of the data. The deviations from this model are used to determine the outliers in the dataset. A good understanding of the given data is often crucial in order to design a simple and accurate model which doesn't overfit the underlying data thus the problem of outlier detection becomes especially challenging, when significant relationships exist among the different data points. Outlier analysis is a major area for research, especially in the area of structural and temporal analysis which handles large volumes of data.

REFERENCES:

1. https://link.springer.com/chapter/10.1007/978-3-319-28549-8_7
2. https://www.knime.com/knime-applications/outlier-detection-in-medical-claims
3. https://www.knime.com/.../anomaly-detection-in-predictive-maintenance-with-time
4. https://nodepit.com › Nodes › Community Nodes › HCS Tools › Pre-Processing
5. https://www.researchgate.net/.../292606298_Outliers_detection_method_using_clustering
6. https://www.scribd.com/document/.../Knime-Anomaly-Detection-Visualization
7. J. Davis, M. Goodrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, in: ICML '06, ACM, 2006.
8. https://dzone.com/articles/outlier-detection-from-large-scale-categorical-bre-1
9. https://pdfs.semanticscholar.org/60f3/49016368f33178e6fd493cd787f8f3f18e01.pdf