International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Outlier Detection by Using: R Software

Dr. Namita Srivastava[1], Ruchi Trivedi[2]

[1]HOD, Department of Statistics, St. Johns College, Agra, India

[2]Research Scholar, Department of Statistics, DR. B. R. Ambedkar University, Agra, India

Abstract: ***Outlier Detection is a critical topic in Data Mining study. The practice of retrieving hidden and usable information from a large data set is known as Data Mining. Data Mining covers authentic, useful and high quality of knowledge. An Outlier is an observation that differs from the rest of dataset's observation. Outlier detection from a collection of datasets is a well-known Data Mining process. Outliers help in detection of unusual patterns and behaviors of different data points which can give a useful result for the research. In data pre-processing and data mining, outlier detection is essential. Outliers have various applications area such as fraud detection, intrusion detection, medical and public health outlier detection, image detection, etc. This paper gives a brief description of outlier's concept its types, causes and applications. Also provides a brief detail about the existing outlier detection algorithm; as to explain the richness and complexity associated with algorithms. Another part of the paper is focused on application of outlier detection algorithm by using R programming. This Software is open statistical source which is use for the analysis of data. In this paper some of the existing algorithms of outlier detection are implemented on our data set by using R Software. Data is taken from R software; the most fascinating part of R is that data is very easily accessible. The Outliers are considered as key in the discovery of unpredicted knowledge.***

Keywords: Outliers, R Software, Outlier detection, Data Mining, Outlier detection methods

1. Introduction

Outlier detection is an aggregation of patterns present in the dataset is an endless problem in the data-mining field. Outliers are that observation that significantly deviates from the other observations in a dataset. Outliers are those data points that cannot be fitted in any type of clusters. Due to its widespread use in a wide range of applications, outlier detection remains a critical and significant research field in data mining. Outlier detection from a collection of datasets is a well-known data mining process. Outliers help in detection of unusual patterns and behaviors of different data points which can give a useful result for the research. In data pre-processing and data mining, outlier detection is essential. Till now, we can assume there is no standard method for detecting outliers since it is very complicated to define characteristics on which outliers can be point out. Outlier detection, now a days become the most researchable area in data mining. This paper comprising of broad literature survey for outliers and different algorithms to detect outliers. Various types of application area of outliers are also explored. The primary goal of outliers is to find objects in large datasets that behave differently than the average object in the data [1]. The rise in data Dimensionality is seen as major hurdle in data mining process.

Dimensionality refers to the number of features or variables. Due to rising size of data many issues such as data redundancy, missing data and so on develop [2]. Outlier detection aims to detect unusual patterns that deviates from the remaining data sets. High dimensional data creates many hinderance for outlier detection, because if the number of variables increases; resulting data becomes sparser in which data points more scattered. As the volume of data grows exponentially, detecting outliers is becoming increasingly difficult. Agarwal. C [3], provides a brief comprehensive study about the outlier analysis comprises its explanation, methods and applications. It combines data mining, machine learning and statistical methodologies with a computational framework, allowing it to be applied to a variety of fields. kaur*et al*. [4], provides a study of several outlier detection algorithm as, well as comparison study of outlier detection methods in order to determine which outlier detection approaches are best suited for high-dimensional data. Outlier detection is important, especially as data pollution grows.

2. Data Mining

Now a days, there is tremendous accumulation of data in the database. Currently, the increment in data is become very rapidly. For this reason, manual data analysis and data retrieval become very time consuming and costly also. The highly growing of data makes a realization, that data mining becomes necessity for data analysis. The practice of retrieving hidden and useable information from a large data set is known as Data Mining. Acquiring knowledge and decision making is a big aspect of data mining. In previous time researchers are managing small scale of data with a smaller number of attributes and a smaller number of records, cannot give exact results. But in today era, as the size and dimension are increasing, still it is not possible to give precise and efficient patterns. This problem gives a direction towards the outlier detection necessity. Data Mining is a multidisciplinary area of computer science and it can be applied with artificial intelligence, statistics and machine learning.

3. Outliers

Outlier detection is an aggregation of patterns present in the dataset is a endless problem in the data-mining field. Outliers are those observations that significantly deviates from the other observations in a dataset. Outliers are those data points that cannot be fitted in any type of clusters. Due to its widespread use in a wide range of applications, outlier detection remains a critical and significant research field in data mining. Till now, we can assume there is no standard

method for detecting outliers since it is very complicated to define characteristics on which outliers can be point out. Outlier detection, now a days become the most researchable area in data mining. This paper comprising of broad literature survey for outliers and different algorithms to detect outliers. Various types of application area of outliers are also explored.

Outliers are the essential area for various discipline such as data mining, statistics, artificial intelligence. Any data point that appears to be out of place in relation to other data points is referred as an Outlier. Outlier detection is the key discovery in many research areas. Mostly outliers are considered as a noise but they both are different; noise is a random error that must be eliminated before outlier detection can be performed. Outlier detection aims to detect unusual patterns that deviates from the remaining data sets. High dimensional data creates many hinderance for outlier detection, because if the number of variables increases; resulting data becomes sparser in which data points more scattered.

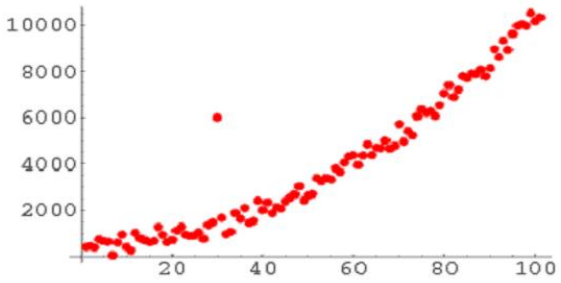

Figure 1: Example of Outlier

a) Causes of Outliers and Reason for Handling Outliers:

Outliers arises due to malicious activity, fraudulent behavior, data entry errors, basically termed as human errors; measurement errors; data extraction error and sampling error (extract data from wrong source). The next issue is the reason behind the study of outliers is that outliers generally considered as errors; but some of them can be important, interested and provides fruitful conclusions. Outliers have a significant impact on dataset. These are the major reasons for the study of outliers.

b) Types of Outliers

A data point is known as a *GLOBALOUTLIER* if any data point is distinct from the complete dataset. It is one of the simplest forms of outlier. For example: credit card fraud detection. When a data datapoint is anomalous with respect to some context (condition) then data point is said to be *CONTEXTUAL OUTLIER*. For example: weight of adult is 60kg may be normal while weight of a child is 60kg is an outlier. *COLLECTIVE OUTLIER* when a group of data point is anomalous from rest of the entire dataset. For example: Human Electrocardiogram output.

c) Application of Outliers

Outlier detection it is a growing area in data mining research. Due to its immense role in various areas. It is not possible to cover all of the application area in single roof, so few of them are discussed according to the recent research and need.

1) Intrusion Detection: It identifies the unauthorized access in computer networks.
2) Credit Card Fraud Detection: Nowadays, Fraud detection becomes a big issue for the banks using a credit card transaction. For example, if anyone card is stolen, suddenly there is a change will be reflecting in purchasing limit of monthly transactions, this change will rise to detect outliers.
3) Fake News and Information, Social Networks: social media becomes a new platform for sharing news, memes and many more things, but it is difficult to verify which news is fake and which one is real. For this, reliable source is used to identify the reality of news; mostly fake news is considered as an outlier.
4) Medical and Public Health Outlier Detection: Many approaches of outlier detection are utilised in medical diagnosis to assist detect critical diseases at early stage and prevent them progressing to a more serious stage.
5) Image Detection: Outlier detection helps to detect the abnormal patterns of image that includes color, texture and brightness.
6) Activity monitoring: For instance, mobile phone fraud can be detected by monitoring phone usage or odd trades in the equity market.

4. Methodology of Research Work

Outlier detection in Data Mining is important topic for research. Outlier Analysis is an essential part of Data Mining in knowledge discovery. The existing methods of outlier detection has been implemented on data with the help of R software. It is an open-source statistical software. For this research work data is collected from R Software which is "Iris Flower dataset which is given by famous biologist Ronald Fisher, it contains 50 samples from each of three species (Iris Setosa, Iris Virginica, Iris Versicolor) and four features also sepal length, sepal width, petal length and petal width. Due to various factors, it has been noted that what appears to be an outlier is actually real time data included in the dataset. For example, in our dataset sudden fluctuation seems in variables sometimes it is increasing and decreasing for some observations, this pattern is known as Outlier. This is now a critical state for identifying this odd pattern (outlier), as it may be led to the discovery of new information and for this Outlier detection methods is implemented on the dataset in R Software for the better results. Different outlier detection methods exist. Few of them like "Statistical outlier detection, Density Based, Distance Based, Clustering Based, Correlation based and Random Forest" have been discussed which is used for detection of outliers.

a) Statistical Outlier Detection

Statistical methods are one of the prior algorithms that can be used for outlier detection. Depth-Based Outlier detection identification is a statistical method that is one of the versions of statistical outlier detection. Outlier data points that are based on depth that are represented by n-d space with assigned depth, data points which have smaller depth is considered as an outlier. Outlier detection is done on the basis of this assigned depth. The data points which have

smaller depth is considered as an outlier. (Tukey, 1977) proposed a graphical tool, "Box Plot", to visualize data [5]. Box plot is main and essential statistical outlier detection technique it works on both type of data multivariate and univariate. Box Plot is a graphical representation of data by quartiles and interquartile which helps in describing the distribution of a data. If any value that is located outside the whiskers of the box plot is considered as outliers means if any value is more than the upper limit or less than the lower limit will be point as OUTLIER. Another renowned statistical approach is GRUBB'S statistics; it works on univariate data; Z value is calculated as the difference between mean value and dubious value is divided by standard deviation to get the Z value. It is also known as extreme studentized deviate test or maximum normalized residual test. Jajo*et al*. [6], suggested an effective adjusted boxplot was introduced to overcome the drawbacks of box-plot for the detection of outlier. Simulationstudy shows the effectiveness of this modification to outlier detection, especially when contamination of data increases.

b) Density Based Outlier Detection

The density of data objects is used in the density-based technique. The density of an object is compared to that of its neighbors in this method. If an object's density is significantly lower than of its neighbors, it is considered to be an outlier. The main principle of the density-based outlier detection algorithm is that in less dense region outlier can be found and in dense regions there will be inliers. Local Outlier Factor is used for detecting outliers. Breuing*et al*. [7] established the concept of density-based clustering based on a comparison with the density of immediate neighborhood. This algorithm requires a prior knowledge about the probability distribution. Local outliers can be detecting efficiently by using this algorithm. Christy *et al*. [8], proposed an approach for detecting outliers that involves that use of two algorithm. The major and important advantages of this approach is that density-based approach is non-parametric approach, they do not rely on any assumed distribution to fit the data. In scalability aspect, they scale better in multidimensional space and in comparison, to statistical methods they are computational efficient. Due to the curse of dimensionality, their performance in high-dimensional space suffers as a result of this method. Using KNN neighborhood along with distance-based approach in high dimensional is being too expensive.

c) Cluster Based Outlier Detection

Clustering based algorithm mainly rely on clusters to define the data behavior. For this smaller size clusters that have lesser data points in comparison to other clusters. Clustering is one of the earliest algorithms for the detection of outliers. Clustering based algorithm is supervised and it will not require any prior knowledge or any trained dataset. This algorithm is quite effective because the data is separated into clusters, this technique is highly applicable. Various clustering algorithms are used for the outlier detection such as Hierarchical clustering which further divides into Agglomerative (bottom-up approach, each data points starts in its own clusters) and Divisive (top-down approach, one cluster is divided into hierarchy) one of the important clustering algorithms is K-means (it works on the principle of distance measure and centroid), K-medoids is also used for outlier detection. Clustering algorithm is robust to different types of datasets. Rajaraman *et al*. [9], proposed an overview of clustering techniques as well as discussion of high-dimensional data. Due to high dimensional data, outlier detection becomes difficult. Outliers are thought to behave in one of two ways; they either do not belong to any cluster or they compelled to belong to a cluster in which they are quite dissimilar to other members, or they belong to other cluster [10]. Behera [11], provides a clustering-based algorithm in low dimensional region as well as high dimensional region. Clustering is widely used approach in many areas for classification, outlier detection and data analysis.

d) Distance based Outlier Detection

The distance between the data points is used in this algorithm. The distance between two points is choosen and then checked. If the neighbouring point are closer, then the situation is regarded as normal; nevertheless, if the neighbouring points are far apart then the situation is considered as Outlier. The distance used for the detection of outliers in this algorithm is k-nearest neighbour (KNN). One of the earliest computing distance studies was given by Knorr and Ng (1999) as: "If at least fraction p of the objects in T lie at a distance greater than D from O, an item O in a dataset T is a DB (p, d) considered as outlier" [12]. Distance-based outlier detection stop for excessive calculation which can be related to fit the observed data points into standard distribution and in choosing outlier test. This algorithm is straightforward and do not depend on any type of distribution to fit the data. Distance based have more robust foundation and more efficient and less efficient complexity in comparison to statistically based method. Distance based have more solid foundationas well as efficient and inefficient complexity in comparison to statistically based method. Various types of distance measures are available to detect outliers. Mostly Euclidean distance is used to detect outliers. But new distances like Hamming distance is used to calculate the distance between two strings. Another distance is cosine it is used for measuring the distance between two vectors. Distance method is efficient approach for low dimensional area, it does not provide best estimates while performing is high dimensional area. In high dimensional data set, number of attributes becomes so large and it does not easy to calculate Euclidean and Manhattan distance for large data set for multiple attributes because it becomes so complex and also it increases computational cost along with time. So, distance measure is more robust for low dimensional region as compare to high dimensional region.

e) Random Forest Outlier Detection

The random forest is like a classifier which is an ensemble learning method. It is another variant of bagging ensemble proposed by Breiman [13]. The aim of random forest is to detect "outliers". This method is based on fact that anomalous points are simpler to isolate than typical points than data is partitioned randomely. Tama*et al*. [14], proposed an efficient random forest method classifierto improve the performance of anomaly detection in IOT network, with parameter setting.

f) Correlation Outlier Detection

Correlation is a statistical measure that is used to examine the relationship between two variables. It is basically defining the linear relationship between the variables. An outlier in correlation analysis is an observation that does not fit in the trend of our data, and would appears to be extreme value. Outliers influence the correlation measure adversely. In some scenario outliers decreases the value of correlation coefficient and weakness the regression measure. But sometimes outliers increase the correlation value and improves the regression measure. Sreevidya *et al.* [15], includes the assumption for data and methods of outlier detection before it is implemented a prior knowledge is necessary about the data.

5. Results

The "Statistical outlier detection, Density Based, Correlation based, Clustering Based and Random Forest" methods of outlier detection have been implemented on Iris Flower dataset which is given by famous biologist Ronald Fisher, it contains 50 samples from each of three species (Iris Setosa, Iris Virginica, Iris Versicolor) and four features also sepal length, sepal width, petal length and petal width. For implementation of these algorithms of outliers in R programming, knowledge of R Codes is very important part because without codes program cannot run and results also not found.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

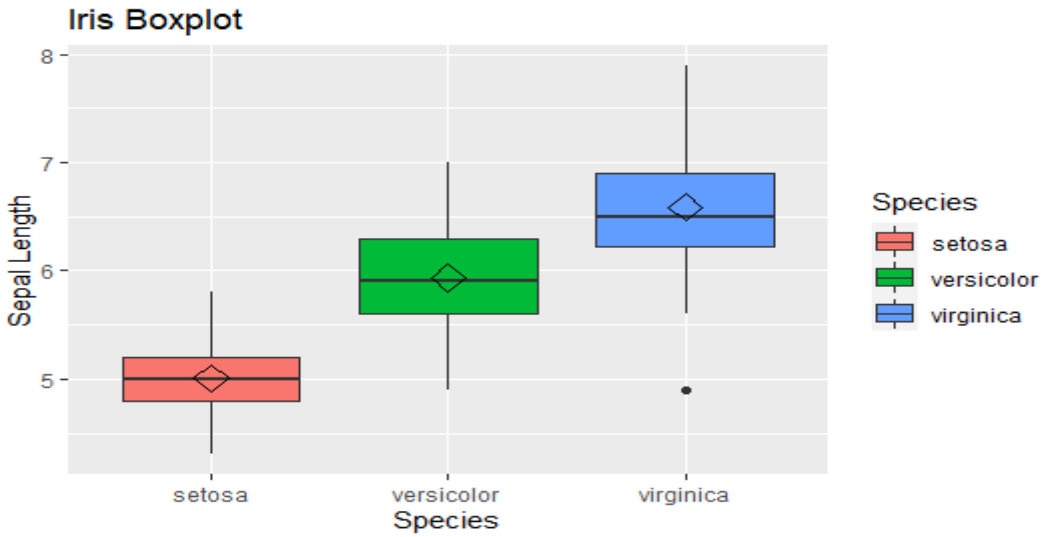

Figure 1: Is showing some observations of given data

Figure 2: Number of Outliers are detected between two features Sepal Length and Species by using Box Plot method. Outliers are those data points which are far away from whiskers plot. So, it is clearly visible in figure

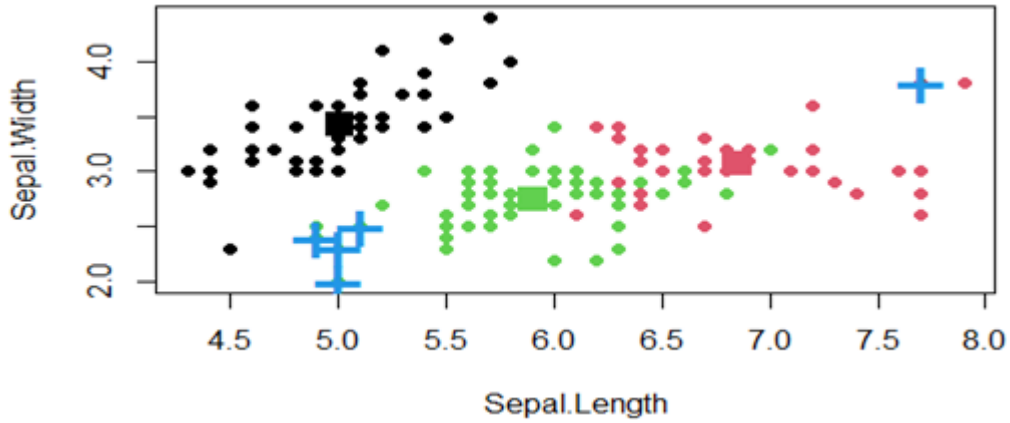

Figure 3: Number of Outliers is detected by using K-Means Clustering Methods for sepal length and sepal width variables of the given dataset. The result shows so many clusters for each variable and the + sign are seems to be outliers because it is not included in cluster or we can say that these points are far away from the clusters.

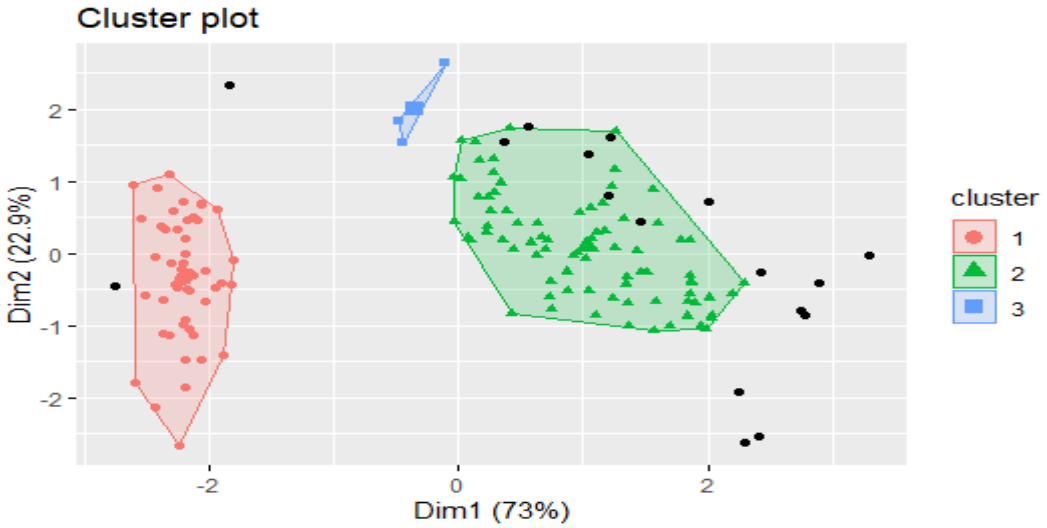

Figure 5: Number of Outliers is detected by using Density-Based ClusteringMethod. The results are shown basis on the density of local neighbourhood points. Some points are seems not to be concentrated in density of cluster, so these points can be considered as Outliers.

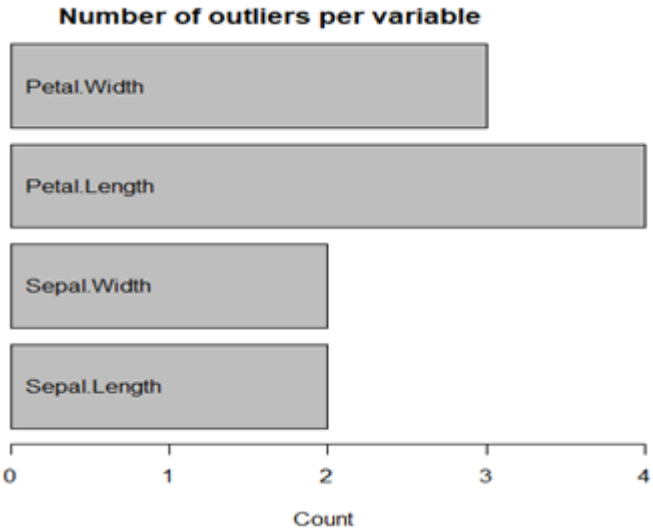

Figure 6: Number of Outliers is detected by using Random-Forest Method. Number of Outliers for each variable are easily visible.

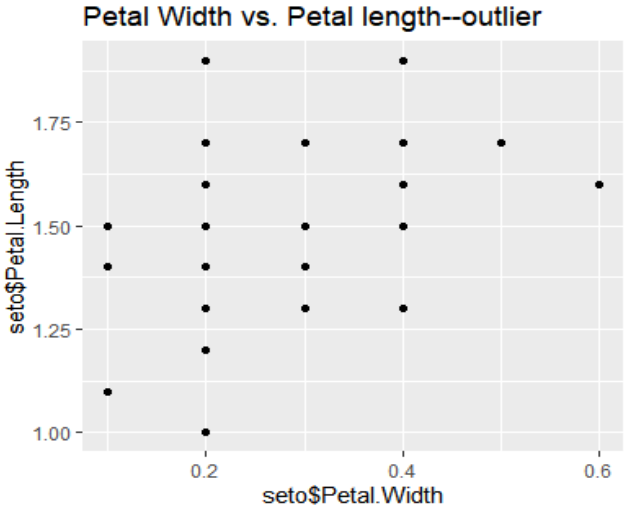

Figure 7: Number of outliers are detected between petal, Width and petal. length of species setosa. Outlier is that point which is far away from all other points in the given figure.

6. Conclusion

We have provided a comprehensive survey in a well-structured manner in this research paper of detecting outliers into a different category by grouping them. Detection of outlier is crucial aspect because it contains useful information which may give a fruitful result for further research. Another focus of the paper is all about a basis of implementing outlier detection methods for detection of outliers for iris flower dataset by using R Software. Data accessibility is easily available in R Software another important point is that knowledge of R codes is essential to implement the algorithms on R software. Based on the study of paper we can conclude that methods for finding outliers and its implementation in R Software. It should be kept in mind that methods of detecting outlier varies by domain to domain. It is important to note that the effectiveness of outlier detection method is largely reliant on the data format. Some of the methods mentioned in this paper necessitate prior knowledge of the data. Selection of different domain is not bounded. Outlier detection approaches gives a straightforward and tangible result for the given data. It is a good work if you want to learn more about the outlier and its different domain. It has been a good work for those who want to start their research on outlier detection and its domain. The whole work is divided into different parts and contains numerous theoretical and practical notions regarding the anomalies.

7. Acknowledgment

I would like to offer my heartfelt appreciation to everyone who helped me produce this research report. This paper and research behind it would not be possible without the extreme support of my supervisor. She has taught me the research methodology and to carry out my research. Throughout the process, I am grateful for their diligent direction, constructive criticism and friendly valuable suggestions. I am grateful to every one of them for sharing their accurate and enlightening knowledge with me.

References

[1] Jayanta K. Dutta, Bonny Banerjee, Chandan K. Reddy, "RODS: Rarity based Outlier Detection in a Sparse Coding Framework", *IEEE Transactions on Knowledge and Data Engineering*, vol.28, issue 2, pp 483-495, September 2015.

[2] R. Lakshmi Devi, Dr. R. Amalraj "An Efficient Unsupervised Cluster based Hubness Technique for Outlier Detection in High dimensional data", *International Journal of Innovative Research in Advanced Engineering*, vol.2, issue 10, pp 63-70, October 2015.

[3] Charu. C. Agarwal, "Outlier Analysis", January 2013.

[4] Kamaljeet Kaur, Atul Garg, "Comparative Study of Outlier Detection Algorithms", *International Journal of Computer Applications* (0975-8887), vol.147-No.9, August 2016.

[5] John Tukey, "An efficient method for displaying a five number data summary", 1997.

[6] NethalJajo, K. M. Matawie, "Outlier Detection using Modified Boxplot", *International Journal of Ecology and Development*, vol.13, pp 116-122, January 2009.

[7] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jorg Sander, "LOF: Identifying density based local outlier", *International Conference on Management of Data*, May 1.

[8] A Christy, G. Meera Gandhi, "Cluster Based Outlier Detection Algorithm for Healthcare Data", Elsevier, vol.50, pp 209-215, 2015.

[9] A. Rajaraman, J. D. Ullman, "Mining of massive datasets", *Cambridge University Press*, Cambridge, 2012.

[10] M. F Jiang, "Two-Phase clustering process for outlier detection", vol.22, no.6-7, pp 691-700, 2001.

[11] H. S. Behera, "A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol.2, issue 4, pp 287-292, April 2012.

[12] Knorr, E. M, Ng, R. T., "Finding International Knowledge of Distance-Based Outliers", *proceedings of the 25th International conference on very large datase*t, Edinburgh, Scotland, pp.211-222, September 1999.

[13] Leo Breiman, "Random Forests", *Machine Learning*, vol.45, pp 5–32, 2001.

[14] RifkiePrimartha, Bayu Adhi Tama, "Anomaly detection using random forest: a performance revised", *International conference on data and Software Engineering*, November 2017.

[15] S SSreevidya, "A Survey on Outlier Detection Methods", *International Journal of Computer Science and Information Technologies*, vol.5 (6), pp 8153-8156, 2014.

[16] Anjali Barmad, Madhu M. Nashipudinath, "An Efficient Strategy to Detect Outlier Transactions", *International Journal of Soft Computing and Engineering*, vol.3, issue 6, pp 174-178, January 2014.

[17] B. Wang, Gang Xiao, Hao Yu, Xiaochun Yang, "Distance-Based Outlier Detection on Uncertain Data", *IEEE International conference on Computer and Information Technology*, vol.1, pp 293 – 298, October 2009.

[18] Bo Liu, Yanshan Xiao, P. S. Yu, Zhifeng Hao, Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels", *IEEE Transactions on Knowledge and Data Engg*, pp 1602 – 1616, 2014.

[19] Charu C. Aggarwal, Philip S, "Outlier Detection for High Dimensional Data", *In proceedings of ACM SIGMOD International conference on Management of data*, vol.30, issue 2, pp 37-46, June 2001.

[20] Dragoljub Pokrajac, Aleksandar Lazarevic, Longin Jan Latecki, "*IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* ", April 2007.