Tribhuvan University

Institute of Science & Technology

Comparative analysis of point outlier detection using cluster based techniques

Dissertation

Submitted To:

Central Department of computer Science & information Technology

Tribhuvan University

Kirtipur, Kathmandu

Nepal

In partial Fulfillment of the requirements for the Degree of Master of science in computer science and information technology

Submitted By:

Kishor Prasad Bhatt

June, 2019

Supervisor

Prof. Dr. Subarna Shakya

Pulchowk Campus, IOE

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Student's Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

… … … … … … …….

Kishor Prasad Bhatt

Date: .....

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Supervisor's Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Kishor Prasad Bhatt** entitled **"Comparative analysis of point outlier detection using cluster based techniques"** be accepted as in fulfilling partial requirement for completion of master Degree of science in computer science and information Technology.

… … … … … … … … … …

Prof. Dr. Subarna Shakya

Department of Electronics & Computer Engineering.

Institute of Engineering,

Pulchowk, Nepal

Date: **……**

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation and in our opinion it is appreciable for the scope and quality as a dissertation in the partial fulfillment for the requirement of Masters Degree in Computer Science and Information Technology.

Evaluation Committee

…… …….

Asst. Prof. Nawaraj Paudel
Head of Department
Central Department of Computer science
& Information Technology
Tribhuvan University
Kirtipur

…….

Prof. Dr. Subarna Shakya
(Supervisor)
Department of Electronics & Computer
Engineering,
Institute of Engineering,
Pulchowk, Nepal

…….
(External Examiner)

…….
(Internal Examiner)

Date: ……

Acknowledgement

I would like to express my sincere thanks to my supervisor **Prof. Dr. Subarna Shakya**, Institute Of Engineering, Pulchowk Campus, Nepal for his support, motivation, suggestions and guidance. His advice was inevitable and with his help I was able to work on my own interested field and complete my dissertation on time.

I am also thankful to **Mr. Nawaraj Poudel**, Head of Department, CDCSIT who has provided all the help and facilities, which I required, for the completion of my dissertation.

Moreover, I would like to express my heartfelt gratitude to all my teachers at Central Department of Computer Science and Information Technology, Tribhuvan University who have imparted knowledge in various subjects.

Last but not the least; I would like to express my thanks to all my friends and my lovely parents for direct and indirect supports for the completion of this thesis.

Kishor Prasad Bhatt

Date:……...

Abstract

Outlier detection is the process of finding peculiar pattern from given set of data. Nowadays, outlier detection is more popular subject in different knowledge domain. Data size is rapidly increases every year there is need to detect outlier in large dataset as early as possible.

In this research, comparison of three different cluster based outlier detection algorithm i.e. K-means with OD, Partition around medoids (PAM) with OD and density based spatial clustering algorithm with noise (DBSCAN) is presented. The main aim of this research is to evaluate their performance of those three different cluster based outlier algorithm for different dataset with different dimension. The dataset used for this research are chosen such way that they are different in size, mainly in terms of number of instances and attributes. When comparing the performance of all three cluster based outlier detection algorithms, in average PAM with OD is found to be better algorithm to detect outlier in most cases with accuracy level 98.30% as well as 92.20% precision, 95.30% recall and 93.43% F-measure value.

Keywords:

Outlier detection, K-means with OD, PAM with OD, DBSCAN.

Table of Contents

List of Figures

Figure **Page**

List of Tables

Table **Page**

List of Abbreviations

Abbreviations	Full Form
OD	Outlier Detection
LDOF	Local Distance Based Outlier Factor
PAM	Partition around Medoid
LOF	Local Outlier Factor
DS	Data Set
DB	Distance Based
IDE	Integrated Development Environment
DBSCAN	Density Based Spatial Clustering Algorithm with noise

CHAPTER 1

1. INTRODUCTION

1.1 Introduction to outlier

Data mining is the technique to analyze and retrieve knowledge from large amount of database and transform it into useful information for future use. Outlier detection is one of the important task in data mining which is actually find out the data object that are deviating from the common expected behaviors. Outlier detection and analysis is sometime known as outlier mining. An outlier is an object that is significantly dissimilar from remaining data object in massive data sets. Hawkins [1] provides formal definition of outlier as following:

"An outlier is an observation which deviates so much from other observation as to arise suspicious that it was generated by different mechanism."

Outlier detection is process of identifying the data object, which drastically different from the rest of data set. Outlier is very important research in data mining field. Outliers occurs due to many reasons like human error, mechanical fault, changes in system behaviors, experimental behaviors etc.[2] Outlier detection is a data mining task with broad applications such as fraud detection, medical diagnosis, image processing , event detection etc.

The outlier can be mainly categorized into three types which are point outlier, contextual outlier and collective outlier. If an individual data point or object can be considered as anomalous with respect to rest of the data, then the data point is called point outlier [3]. If data point is a rare occurrence with respect to some specific context and it is a normal occurrence with respect to some another context such type of data point is called contextual outliers. If a particular data point is not anomalous but it’s with entire data set is anomalous, then it is called collective outliers.

The point outlier is simplest forms of outlier and is the key focus of majority of research in outlier detection [3]**.** Proximity based outlier detection method is one the best approach to detect the point outliers. The proximity outlier detection method can decomposed into three classes namely distance based, density based and cluster based techniques.

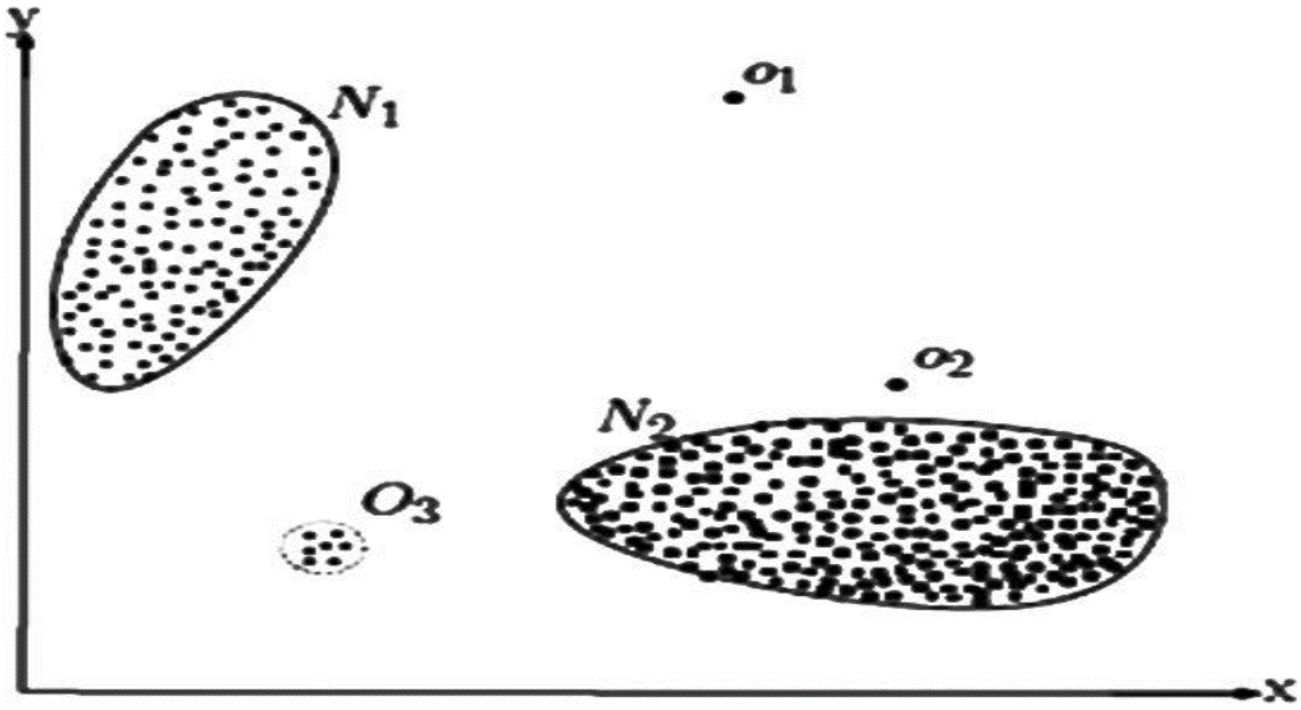

Figure-1: Cluster with outlier.

1.2 Cluster based approach

Clustering based outlier detection is the one of the popular point outlier detection technique. The goal of clustering is that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized [4].

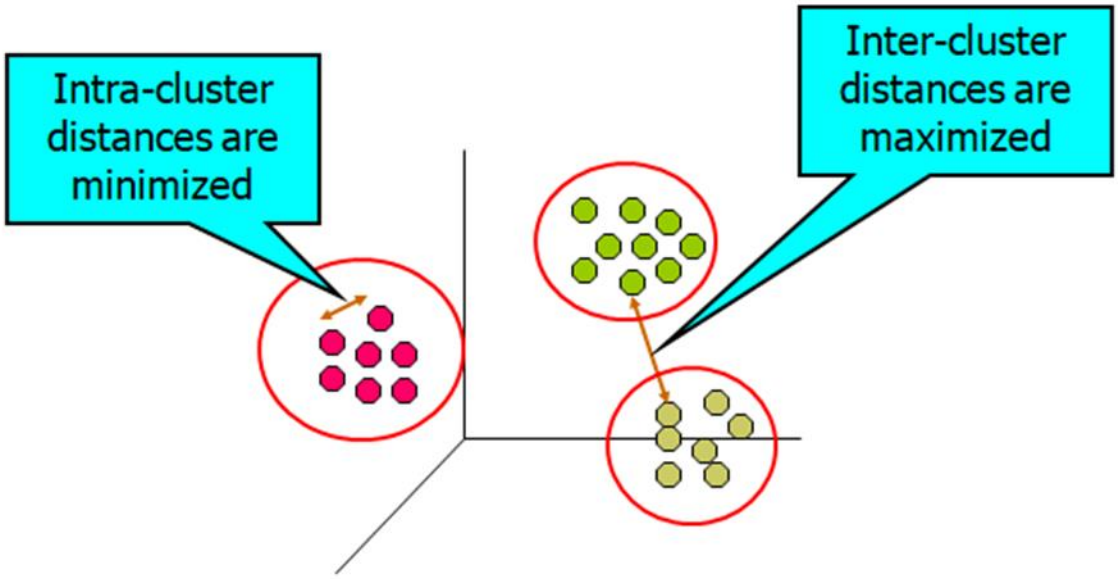

Figure-2: Basic principle of clustering.

This means that object within cluster are as similar as possible and dissimilar as possible with other clusters object. It is a useful technique for the discovery of data distribution and patterns in the original data. The goal of clustering technique is to find out both the dense and the sparse region in a data set. It is a method of unsupervised learning and a common technique for statistical data analysis. It is an important technique used for outlier analysis. Clustering based techniques can find the single point as outlier or whole cluster as outlier depends on the criteria.

1.3 Problem statement

Data mining application has got rich focus due to its significance of outlier detection algorithms. The comparison of outlier detection algorithm is complex and open problem. The point outlier detection from given data set is crucial task in data mining nowadays. There are various traditional algorithms for detecting point outlier from the dataset and most of them are vulnerable to detect outliers. So clustering based approach is one of the widely used approach to detect point outliers. There are different algorithms available to detect the point outliers. The selection best clustering algorithm for given dataset is widespread problem. So algorithm performance measure parameter like accuracy, precision, recall and F-measure can be used for help in selection of best clustering based outlier detection algorithms for given data set.

1.4 Objective of thesis

The main objectives of this research are:

- To detect point outliers from given data set by using cluster based outlier detection algorithms (K-means with OD, PAM with OD and DBSCAN).
- To perform comparative analysis of these clustering based outlier detection algorithms based on accuracy, precision, recall and F-measure parameters.

1.5 Limitation of thesis

Limitations of this research were:

- This study had done by comparison between three algorithms of cluster based outlier detection algorithms. (K-means with OD, PAM with OD, DBSCAN)
- This research had focused on comparison of Accuracy, Precision, Recall, and F-measure of implemented algorithms.
- All algorithms had implemented in R-programming language.

1.6 Structure of Report

This report is organized in six chapters including the following chapters.

- Chapter 1 of this dissertation work is introduction part, which is organized into subsequent five chapters.
    - First chapter is focused on introduction and overview of outlier and
    - Second chapter is about Cluster based outlier approach.
    - Third chapter is about problem analysis of existing or previous works which demands further study to get better solutions.
    - Fourth chapter describe the main objective of this dissertation works.
    - Fifth chapter is about limitation of this dissertation works.
- Chapter 2 contains explanation of all previous studies related to this topic in detail under literature review.
- Chapter 3 includes details of all algorithms studied.
- Chapter 4 describes the implementation details.
- Chapter 5 contains all the details of data which is applied for analysis purpose and comparative performance measure of all three different proximity based outlier detection

algorithms over three different datasets. The result of the study is shown in tabular form as well as in graph.

- Chapter 6 provides final conclusion and future works of the study.

CHAPTER 2

2. LITERATURE REVIEW

2.1 Background and Literature Review

Outlier detection is very essential of any modeling techniques. A failure to detect outliers or their ineffective handling can have serious effects on the strength of the inferences drained from the technique. There is large number of techniques available to perform this task, and often selection of the most suitable technique poses a big challenge to the practitioner. Therefore many approaches have been developed to detect outliers. There are several factors that determine how to formulate an outlier detection problem.

Statistical methods are one of the earliest algorithms that can be used by various outlier detection methodologies [5]. One of the single dimensional unilabiate methods is Grubb's method which is Extreme Studentzed Deviate [6] which calculates a z value as the difference between the mean value of the attribute and the query value divided by standard deviation for the attributes.

Anguilli and Pizzuti[7] proposed a distance based method to determine the outliers by considering the whole neighborhood of the objects. All the points are ranked based on the sum of distances from the k-nearest neighbors.

Breuing et.al [8] proposed a Local Outlier Factor for each each object in the data set, indicating its degree of outlierness. This is the first concept of density based outlier detection algorithms. Since LOF value of each object is obtained by comparing its density with those in its neighborhood.

In [9] proposes a method based on clustering approaches for outlier detection. They first perform the clustering algorithm on data, after performing the clustering then different outlier detection methods are used.

In [16] an algorithm that provides OD and data clustering simultaneously. The algorithm improve the estimation of centroids of the generative distribution during the process of clustering and outlier discovery. The first stage consists of IGK process, while second stage iteratively removes the vectors which are far from their cluster centroids.

In [15] to detect outliers includes three methods which are clustering, pruning and computing outlier factor. Clustering is used which partition the dataset into given number of clusters. In

pruning, based on distance measure, points which are closed to the centroid of each cluster are pruned. For unpruned points, LDOF measure is calculated, a measure LDOF , tells how much data deviating from its neighbours. The high LDOF value of points indicates that the point may be outlier.

In [10] perform the K-means clustering and then calculate the mean of all data points, after this compare the distance of all points to the mean and if distance is greater than mean, those points are considered to be outliers.

In paper [2] perform the k-means clustering, then calculates the threshold (T), all points grater then T distances are considered as outlier and find small cluster for some specified criteria and all points in the cluster are considered as outlier points.

In [11] perform the PAM clustering and then first find small cluster with specified criteria, all object in the cluster are considered as outlier and secondly calculates the distance of all objects with respect to centroid and choose the K farthest points and those points are considered as outlier points. In this study perform the PAM clustering, then calculates the threshold (T), all points grater then T distances are considered as outlier and find small cluster for some specified criteria and all points in the cluster are considered as outlier points.

In [12] all clustering algorithm find the circle shaped cluster, but real world data is not always such. So M. Ester et al introduce the DBSCAN algorithm which can find the arbitrary shaped cluster with outliers.

In [13] the DBSCAN algorithm is used to find outlier in temperature dataset and results that it works better then statistical approach.

CHAPTER 3

3. RESEARCH METHODOLOGY

Clustering based outlier detection techniques define a data point as outlier. There are different ways to define the data point as outliers. The entire study is performed as follows.

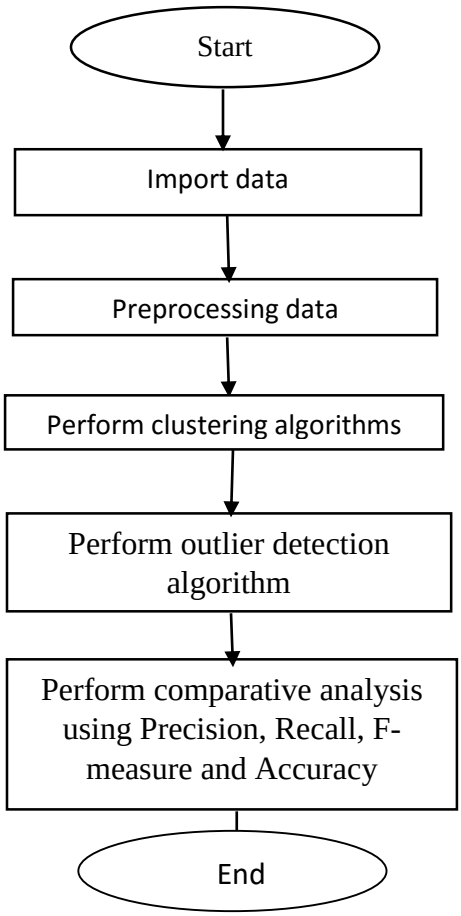

Figure-3: Flowchart of entire process of study.

3.1 Data collection

There are three different type of data set which are collected from UCI machine learning repository. The data set have been chosen such that they are differ in size, mainly in terms of number of instances and number of attributes.

3.1.1 Dataset 1

The first dataset is small Iris dataset. The dataset contains 4 attributes apart from the class attributes with 150 instances. But in this research only 112 instances have been taken in which 100 instances are normal and 12 instances taken as outlier.

3.1.2 Dataset 2

The second dataset is medium sized Seed dataset. The dataset contains 7 attributes apart from the class attributes with 210 instances. But in this research only 162 instances have been taken in which 140 instances taken as normal and 22 instances taken as outlier.

3.1.3 Dataset 3

The large data set is Breast Cancer Wisconsin (Diagnostic). The dataset contains 30 attributes apart from the id and class attributes with 569 instances. But in this research only 397 instances have been taken in which 357 instances taken as normal data and 40 instances taken as outlier data.

3.2 Tool Used

All the algorithms are implemented in R-language using R x64 3.4.1 with use of R-packages libraries.

3.2.1 Programming language

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. R made its first appearance in 1993. A large group of individuals has contributed to R by sending code and bug reports. Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive. As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

- ✓ R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- ✓ R has an effective data handling and storage facility,
- ✓ R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- ✓ R provides a large, coherent and integrated collection of tools for data analysis.
- ✓ R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

As a conclusion, R is world's most widely used statistics programming language. It's the choice of data scientists and supported by a vibrant and talented community of contributors

3.2.2 RStudio

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, Red Hat/Centos, and SUSE Linux). Using RStudio we can perform syntax highlighting, code completion, and smart indentation, execute R code directly from the source editor, quickly jump to function definitions, easily manage multiple working directories using projects, integrated R help in documentation, interactive debugger to diagnose and fix errors quickly.

3.3 Cluster based algorithms

Clustering based outlier detection techniques define a data point as outlier. There are different ways to define the data point as outliers. The most common clustering based outlier detection and analysis techniques are as follows:

3.3.1 K-means algorithm [2]

K-means is most popular clustering algorithm. The main objective is to partitioning n objects into k clusters, so that the inter cluster similarity is minimum and intra cluster similarity is maximum.

Input:

K: no. of cluster

D: dataset containing n objects

Output: A set of k clusters.

Method:

1. Arbitrarily choose k objects in D as the initial cluster Centre.
2. Calculate the distance between each data point and cluster centers using distance measure formula.

$$\text{Distance (d)} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

3. Assign the data point to the cluster centre whose distance from the cluster centre is the minimum of all the cluster centres.

4. When all the objects are placed recalculate the centroid k position using following formula.

$$\text{Mean } (c_i) = \frac{1}{k_i}\sum_{x_i \epsilon k_i} x_i$$

5. Repeat step 2 and 3 until position of k is no longer moved.

After clustering, firstly we have to check each cluster with number of element less than the one third of (n/k) then all elements in the cluster are considered as an outliers. Secondly, calculate the threshold value as follows:

1. Calculate the pairwise distance for whole data object with respect to cluster centroids using Euclidean distance measure formula.

2. Calculate the Maximum and Minimum value of each cluster and then store the maximum value into an Array $D_{k\text{-}max}$ and store the minimum value in $D_{k\text{-}min}$ Array.

$$D_{k\text{-}max}=\text{Max } (d1,d2,\ldots\ldots,dn)$$

$$D_{k\text{-}min}=\text{Min } (d1, d2,\ldots\ldots,dn)$$

Where k=1,2…,n is the cluster number .

3. The threshold value for each cluster K(1,2,…,n) can be calculated and then the results stored into an array Th using following equation

$$Th_{k\text{-}critical} = (D_{k\text{-}max} + D_{k\text{-}min})/2$$

where k is the cluster number.

3.3.1.1 Outlier detection algorithm

Input: k Clusters of n data (Output from K-mean Algorithm)

Output: Set of outliers

1. Set threshold value $Th_{k\text{-}critical}$
2. For each cluster **k**

2.1 If total number of elements is less than one third of (n/k), add all the elements into Outlier list

2.2 Else, calculate distance from cluster centre to all the data set in the cluster.

If distance > threshold value ($Th_{k\text{-}critical}$) this data is considered as outlier.

Else

Data is not outlier.

3.3.2 PAM Algorithm [11]

PAM algorithms is a partition based clustering algorithm. In this algorithm, partition the n data object into the k number of cluster based on the closest cluster center (mediod).

Input: Set of n data.

Output: Set of k clusters containing all of n data in any of the clusters

1. Initialize: randomly select (without replacement) k of the n data points as the medoids
2. Associate each data point to the closest medoid. ("closest" here is defined using Manhattan distance formula)

$$\text{Manhattan distance (d)} = \sum_{i=1}^{n} |x_i - y_i|$$

3. For each medoid m

   i) For each non-medoid data point o

   ii) Swap m and o and compute the total cost of the configuration using following formula.

$$\text{Total cost } (C_i) = \sum_{i=1}^{n} |E_i - O_i|$$

4. Select the configuration with the lowest cost.
5. Repeat steps 2 to 4 until there is no change in the medoid.

After clustering, firstly we have to check each cluster with number of element less than the one third of (n/k) then all elements in the cluster are considered as an outliers. Secondly, calculate the threshold value as follows:

1. Calculate the pairwise distance for whole data object with respect to cluster centroids using Euclidean distance measure formula.

2. Calculate the Maximum and Minimum value of each cluster and then store the maximum value into an Array $D_{k\text{-}max}$ and store the minimum value in $D_{k\text{-}min}$ Array.

$$D_{k\text{-}max}=\text{Max } (d1,d2,\ldots\ldots,dn)$$

$$D_{k\text{-}min}=\text{Min } (d1, d2,\ldots\ldots,dn)$$

Where k=1,2…,n is the cluster number .

3. The threshold value for each cluster K(1,2,…,n) can be calculated and then the results stored into an array Th using following equation

$$Th_{k\text{-}critical} = (D_{k\text{-}max} + D_{k\text{-}min})/2$$

where k is the cluster number.

3.3.2.1 Outlier detection algorithm

Input: k Clusters of n data (Output from PAM Algorithm)

Output: Set of outliers

3. Set threshold value $Th_{k\text{-}critical}$
4. For each cluster **k**

2.3 If total number of elements is less than one third of (n/k), add all the elements into Outlier list

2.4 Else, calculate distance from cluster centre to all the data set in the cluster.

If distance > threshold value ($Th_{k\text{-}critical}$) this data is considered as outlier.

Else

Data is not outlier.

3.3.3 DBSCAN algorithm

DBSCAN is a density based clustering technique. Density-based algorithms like DBSCAN find the core objects at first and they are growing the clusters based on these cores and by searching for objects that are in a neighborhood within a radius epsilon of a given object. The advantage of these types of algorithms is that they can detect arbitrary form of clusters and it can filter out the outliers. [14] Those points are considered to be outliers which are not placed in any cluster.

Algorithm

Input:

D: the dataset

Eps: the neighborhood distance

Minpts: the minimum number of points

Output:

Discovered outliers and clusters.

The main steps of DBSCAN algorithm are as follows:

- Arbitrarily select a point P.
- Retrieve all points density-reachable from P with respect to Eps and Minpts.
- If P is a core point, a new cluster is formed or existing cluster is extended.
- If P is a border point, no points are density-reachable from P, and DBSCAN visits the next point of the database.
- Continue the process with other points in the database until all of the points have been processed.
- DBSCAN may merge two clusters into one cluster, if two clusters of different density are close to each other. They are close if the distance between clusters is lower than Eps.

3.4 Comparison Criteria

The comparative analysis or the result is made on the basis of following criteria.

3.4.1 Confusion Matrix:

A confusion matrix is a table for analyzing the result of outlier classifier. It deals with how outlier detection algorithm can recognize tuples of different outlier class (Either yes or no). in order to develop the confusion matrix , the following terms should be considered[17].

- **True Positive (TP):** Positive tuples that are correctly labeled by the outlier detection algorithm.
- **True Negative (TN):** Negative tuples that are correctly labeled by outlier detection algorithm.
- **False Positive (FP):** Negative tuples that are incorrectly labeled as positive.
- **False Negative (FN):** Positive tuples that are mislabeled as negative.

	Predicted outlier	Predicted Normal
Actual outlier	**TP**	**FN**
Actual Normal	**FP**	**TN**

Table 3-1: Confusion Matrix

3.4.2 Accuracy

Accuracy of outlier detection algorithm on given dataset is percentage of dataset tuples that are correctly classified as outlier or not. It also refers to the recognition rate of the outlier detection algorithm.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

3.4.3. Precision

Precision refers to the measure of exactness that means what percentage of tuples labeled as positive (or outlier) are actually such.

$$\text{Precision} = \frac{TP}{TP+FP}$$

3.4.4 Recall

Recall refers to the true positive or outlier that means the proportion of positive tuples that are correctly identified.

$$\text{Recall} = \frac{TP}{TP+FN}$$

3.4.5 F- Measure

The F-measure or F-score also refers to F-measures that combines both measures precision and recall as the harmonic mean.

$$\text{F-measure} = \frac{2 \text{ x precision x recall}}{\text{precsion+recall}}$$

CHAPTER 4

4. RESULT, ANALYSIS AND COMPARISONS

4.1 Result Analysis and comparison

In this study, the analysis of all three algorithms mentioned in chapter 3 is compared for three different dimensional dataset mentions in chapter 3.1 which is compared based on accuracy, precision, recall and F-measure. The results were achieved by using whole test dataset for different outlier algorithms.

4.1.1 Comparison results of proximity-based outlier detection algorithms for dataset1

Table 4-1 provides the summery output for comparison of all three algorithms studied over dataset (i.e. Iris dataset). In this dataset total 112 observations are taken, 100 observations are considered as normal and 12 observations considered as an outlier.

Algorithms	Accuracy	Precision	Recall	F- measure
K-means with OD	91.90%	57.10%	100%	72.70%
PAM with OD	98%	85.70%	100%	92%
DBSCAN	90%	52.10%	100%	68.60%

Table 4-1: Result of outlier Detection for Data set 1.

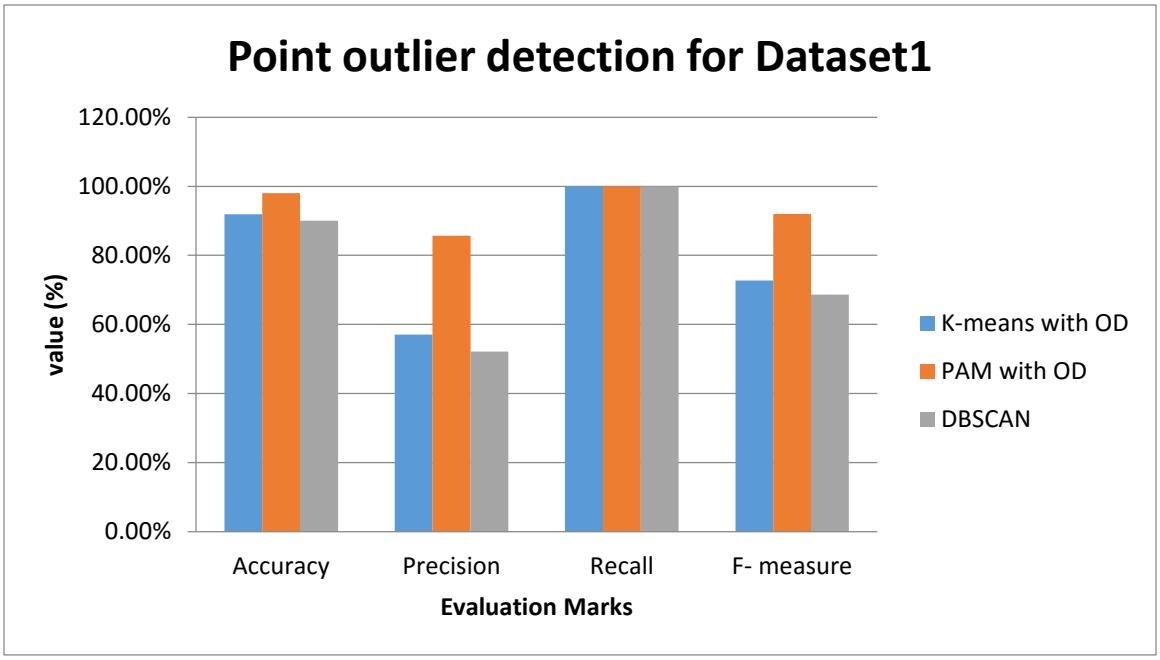

Figure 4-1: Graph of table 1

Based on the Figure 4-1, it is clearly seen that the accuracy value of PAM with OD is got high label of 98% and DBSCAN got less accuracy of label 90%. In case of precision and the Recall value of implemented cluster based outlier detection algorithms PAM with OD had got high precision and recall label of 85.70% and 100% respectively. Whereas DBSCAN got less precision and recall level of 52.10% and 90 % respectively.

Figure 4-1 also show the F-measure of table 4- observed by implemented cluster based outlier detection algorithms. Again PAM with OD had got a victory over compared algorithms with value of 92% and DBSCAN had got minimum value of 68.60%.

4.1.2 Comparison results of proximity based outlier detection algorithms for dataset 2:

Table 4-2 provides the summary output for comparison of all three algorithms studied over dataset (i.e. Seed dataset). In this dataset total 162 observations are taken, 140 observations are considered as normal and 22 observations considered as an outlier.

Algorithms	Accuracy	Precision	Recall	F- measure
K-means with OD	96.90%	94.70%	81.80%	87.70%
PAM with OD	97.50%	90.90%	90.90%	90.90%
DBSCAN	96.20%	80.70%	95%	87.40%

.

Table 4-2 Result of outlier detection algorithm for dataset 2

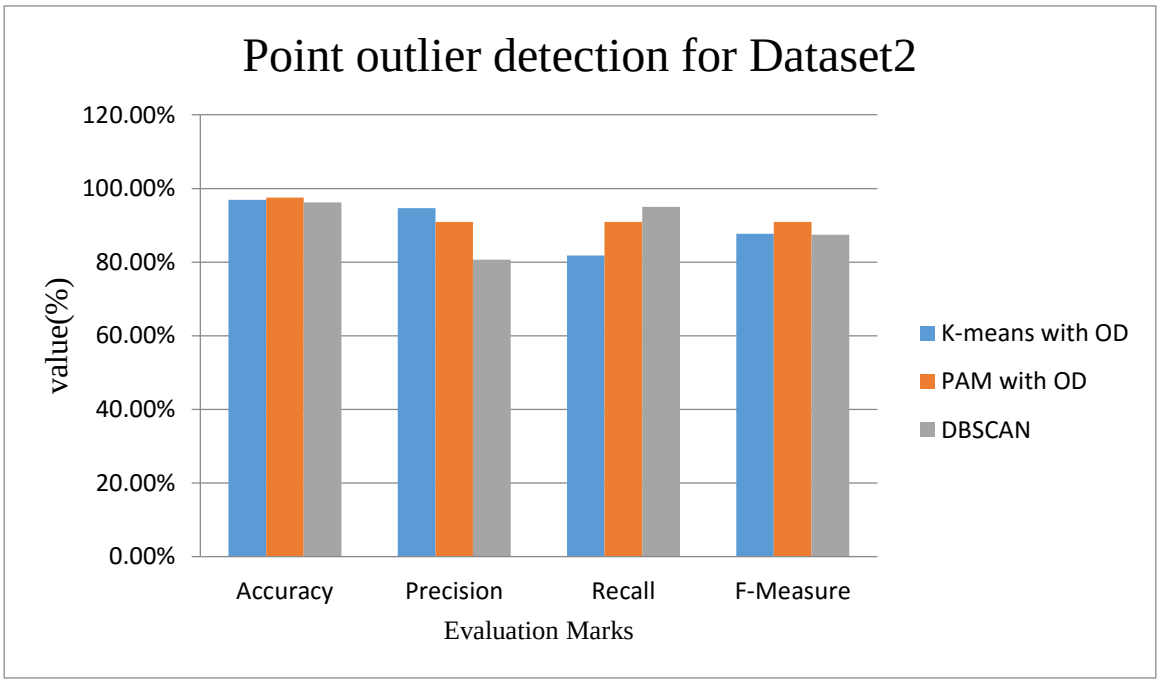

Figure 4-2: Graph of table 4-2.

Based on the Figure 4-2, it is clearly seen that the accuracy value of PAM with OD is got high label of 97.50% and DBSCAN got less accuracy of label 96.20%. In case of precision implemented cluster based outlier detection algorithms K-means with OD had got high precision 94.70%. Whereas DBSCAN got less precision label of 80.70%. In case of recall DBSCAN has got high label of 95% and K-means with OD got less label of 81.80%.

Figure 4-1 also show the F-measure of table 4-2 observed by implemented cluster based outlier detection algorithms. Again PAM with OD had got a victory over compared algorithms with value of 90.90% and DBSCAN had got minimum value of 87.40%

4.1.3 Comparison results of proximity based outlier detection algorithms for dataset 3

Table 4-2 provides the summary output for comparison of all three algorithms studied over dataset (i.e. Breast cancer dataset). In this dataset total 397 observations are taken, whereas 357 observations are considered as normal and 40 observations considered as an outlier.

Algorithms	Accuracy	Precision	Recall	F- measure
K-means with OD	95.40%	100%	55%	70.90%
PAM with OD	99.4%	100%	95%	97.40%
DBSCAN	96%	79.10%	95%	86.30%

Table 4-3: Result of outlier Detection for Data set 3.

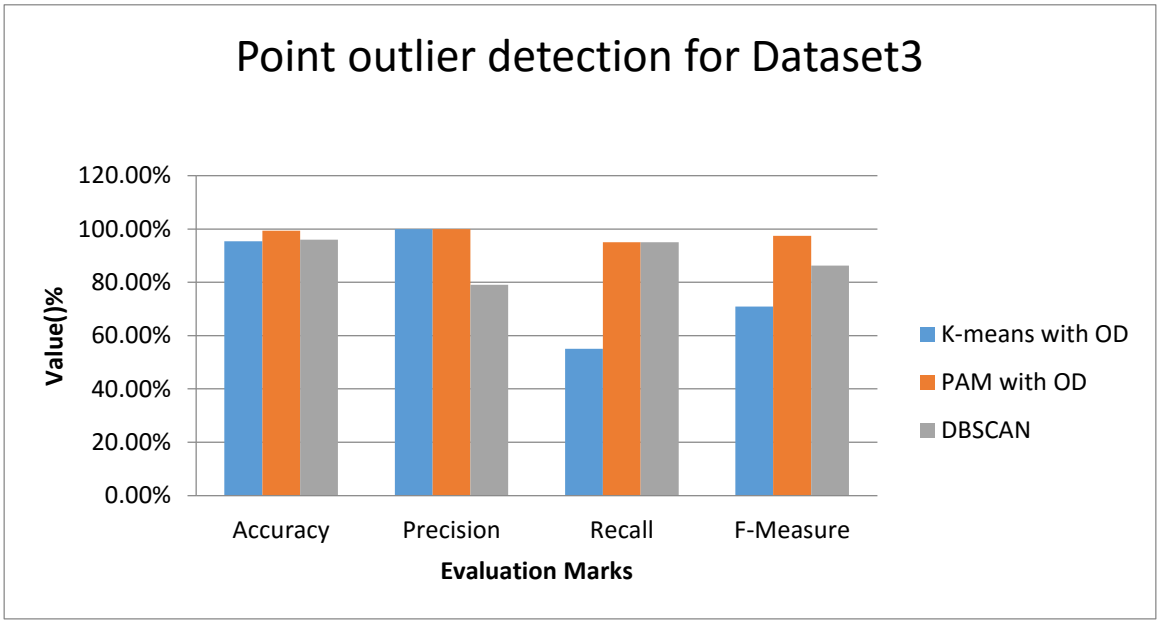

Figure 4-3: Graph of table 4-3

Based on the Figure 4-1, it is clearly seen that the accuracy value of PAM with OD is got high label of 99.40% and K-means with OD got less accuracy of label 95.40%. In case of precision implemented cluster based outlier detection algorithms PAM with OD and K-means with OD had got high precision of 100%. Whereas DBSCAN got less precision label of 79.10%. In case of

recall, PAM with OD and DBSCAN got high value with 95% whereas K-means with OD got less value that is 55%.

Figure 4-3 also show the F-measure of table 4-3 observed by implemented cluster based outlier detection algorithms. Again PAM with OD had got a victory over compared algorithms with value of 97.40% and K-mean with OD had got minimum value of 70.90%.

4.1.4 Comparison of average results of proximity based outlier detection algorithms:

Table 4-3 provides the summery of average output for comparison of all three algorithms studied over different three dataset (i.e. Iris dataset, seed dataset, Breast cancer dataset).

Algorithms	Accuracy	Precision	Recall	F- measure
K-means with OD	94.73%	83.93%	78.93%	77.10%
PAM with OD	98.30%	92.20%	95.30%	93.43%
DBSCAN	94.07%	70.63%	93.67%	80.76%

Table 4-4: Averages result of outlier Detection for all dataset.

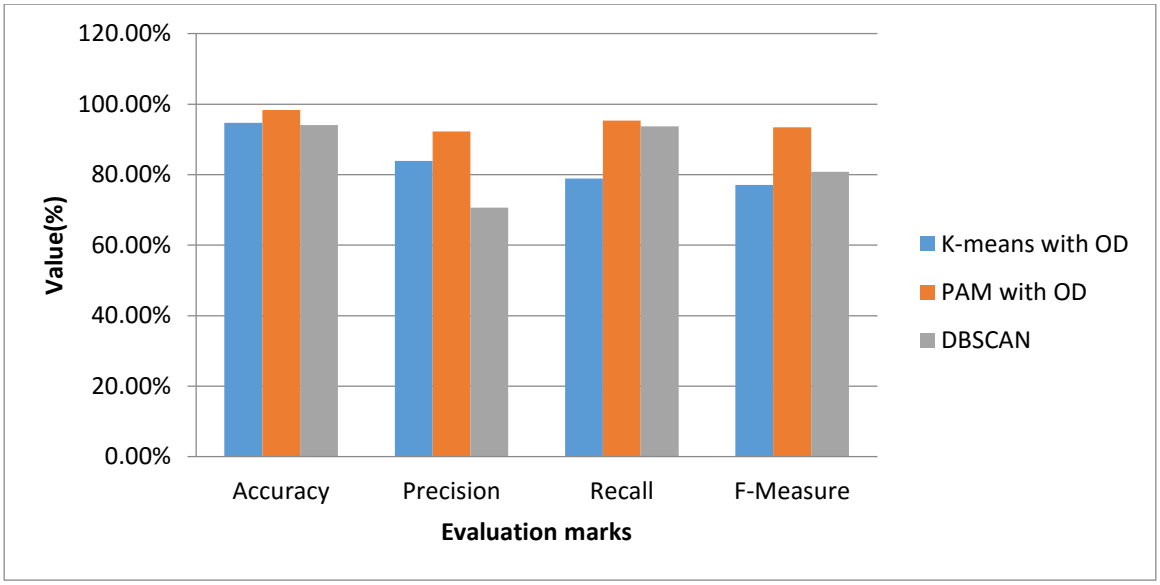

Figure 4-4: Graph of table 4-4

Figure 4-5 showed that comparisons between averages of all evaluation metrics of all implemented cluster based outlier detection algorithms. From that comparison, PAM with OD had got rich as well as motivating and encouraging performance in every aspect, whereas K-means with and DBSCAN had got less performance in every aspect as compared to PAM with OD algorithm.

CHAPTER 5

5. CONCLUSION

5.1. Conclusion

In this research, the comparative analysis of cluster based outlier detection algorithms (i.e. partition based algorithms K-means with OD, PAM with OD and density based spatial clustering with noise(DBSCAN)) using various measure parameter like accuracy, precision, recall and F-measures over the three different dataset(i.e. Iris dataset, seed dataset, breast cancer dataset) with different dimension and size are evaluated. From the result analysis, PAM with OD has higher accuracy as well as higher precision, recall and F-measure with level of 98.73%, 92.20, 95.30% and 93.43% respectively on average as compared to the PAM with OD and DBSCAN.

On balance scale, PAM with OD algorithm has predicted better outlier result than other cluster based outlier detection algorithms studied for all dataset.

More algorithms from the cluster based outlier detection can be incorporated for further study to be studied dataset or other dataset which have numeric as well as categorical value. Moreover some algorithms can be customized for the specific domain so that outlier detection could have more accurate and reliable.

References

[1] Hawkins D. “Identification of Outliers”, *Chapman and Hall*, 1980.

[2]A. Barai and L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering," *World Journal of Computer Application and Technology*, pp. 24-29, 2017.

[3] Malik K., H.Sadawart Kalra G.S. “Comparative Analysis of Outlier Detection Techniques” *International Journal of Computer Applications* (0975 – 8887), vol. 97– no.8, 2014

[4] S. K. Priyanka W. Meshram, "Literature Survey on Outlier Detection Techniques For Imperfect Data Labels," *International Journal of Science and Research (IJSR),* vol. 4, no. 1, pp. 2731-2735, 2013

[5] Barnett, V. and Lewis,“Outliers in Statistical Data”. John Wiley & Sons.3rd edition, 1994.

[6] Huber P. “Robust Statistics”. Wiley, New York, 1974.

[7] Angiulli F., Basta S., and Pizzuti. C., “Distance-based detection and prediction of outliers”*IEEE Transactions on Knowledge and Data Engineering*, pp. 145–160, 2006.

[8] Breuning M., Kriegel H-P., R. Ng, and Sander J., “LOF: Identifying Density-Based Local Outliers. In Proc. of 2000 ACM SIGMOD”*International Conference on Management of Data*, pp. 93-104, 2000.

[9] Al-Zoubi, M.“An Effective Clustering-Based Approach for Outlier Detection”*European Journal of Scientific Research*, 2009

[10]Aggarwal and J. Singh, "Outlier Detection Using K-Mean and Hybrid Distance Technique on Multi-Dimensional Data Set," *International Journal of Advanced Research in Computer Engineering &Technology (IJARCET),* vol. 2, no. 9, pp. 2626-2631, 2013

[11]V. Kumar, S. Kumar and A. K. Singh, "Outlier Detection: A Clustering-Based Approach," *International Journal of Science and Modern Engineering (IJISME),* vol. 1, no. 7, pp. 16-19, 2013

[12]Martin E. et. al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," *International Conference on Knowledge Discovery and Data Mining,* p. 226–231, 1996

[13] M. ÇELİK and e. al., "Anomaly Detection in Temperature Data Using DBSCAN Algorithm," *IEEE International Symposium on INnovations in Intelligent SysTems and Applications,* 2011

[14] Mohamad T., Elbatta and Ashour W., "A Dynamic Method for Discovering Density Varied Clusters," *International Journal of Signal Processing, Image Processing and Pattern Recognition,* vol. 6, no. 1, pp. 123-134, 2013

[15] Pranjali Kasture, Jayant Godge, "Cluster based Outlier Detection" *International Journal of Computer Applications,* Volume 58-No. 10, November 2012.

[16]M. H. Marghny, Ahmed I. Taloba "Outlier Detection Using Improved Genetic K-Means", *International Journal of Computer Applications,* Volume 28-No. 11, August 2011

[17]Sofia Visa et.al. "Confusion Matrix-based Feature Selection" Conference Paper · January 2011.

[18] Emmanuel Paradis, "R for Beginners" At: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

APPENDIX

(Sample Dataset)

1 Sample data for dataset1 (IRIS dataset)

Att1	Att2	Att3	Att4
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1
5.4	3.7	1.5	0.2
4.8	3.4	1.6	0.2
4.8	3	1.4	0.1
4.3	3	1.1	0.1
5.8	4	1.2	0.2
5.7	4.4	1.5	0.4
5.4	3.9	1.3	0.4
5.1	3.5	1.4	0.3
5.7	3.8	1.7	0.3
5.1	3.8	1.5	0.3
5.4	3.4	1.7	0.2
5.1	3.7	1.5	0.4
4.6	3.6	1	0.2
5.1	3.3	1.7	0.5
4.8	3.4	1.9	0.2
5	3	1.6	0.2
5	3.4	1.6	0.4
5.2	3.5	1.5	0.2
5.2	3.4	1.4	0.2
4.7	3.2	1.6	0.2
4.8	3.1	1.6	0.2
5.4	3.4	1.5	0.4
5.2	4.1	1.5	0.1
5.5	4.2	1.4	0.2
4.9	3.1	1.5	0.1
5	3.2	1.2	0.2

5.5	3.5	1.3	0.2
4.9	3.1	1.5	0.1
4.4	3	1.3	0.2
5.1	3.4	1.5	0.2
5	3.5	1.3	0.3
4.5	2.3	1.3	0.3
4.4	3.2	1.3	0.2
5	3.5	1.6	0.6
5.1	3.8	1.9	0.4
4.8	3	1.4	0.3
5.1	3.8	1.6	0.2
4.6	3.2	1.4	0.2
5.3	3.7	1.5	0.2
5	3.3	1.4	0.2
7.1	3	5.9	2.1
6.3	2.9	5.6	1.8
6.5	3	5.8	2.2
7.6	3	6.6	2.1
4.9	2.5	4.5	1.7
7.3	2.9	6.3	1.8
6.7	2.5	5.8	1.8
7.2	3.6	6.1	2.5
6.5	3.2	5.1	2
6.4	2.7	5.3	1.9

2 Sample data for dataset2 (Seed Data)

att1	att2	att3	att4	att5	att6	att7
15.26	14.84	0.871	5.763	3.312	2.221	5.22
14.88	14.57	0.8811	5.554	3.333	1.018	4.956
14.29	14.09	0.905	5.291	3.337	2.699	4.825
13.84	13.94	0.8955	5.324	3.379	2.259	4.805
16.14	14.99	0.9034	5.658	3.562	1.355	5.175
14.38	14.21	0.8951	5.386	3.312	2.462	4.956
14.69	14.49	0.8799	5.563	3.259	3.586	5.219
14.11	14.1	0.8911	5.42	3.302	2.7	5
16.63	15.46	0.8747	6.053	3.465	2.04	5.877
16.44	15.25	0.888	5.884	3.505	1.969	5.533
15.26	14.85	0.8696	5.714	3.242	4.543	5.314
14.03	14.16	0.8796	5.438	3.201	1.717	5.001
13.89	14.02	0.888	5.439	3.199	3.986	4.738

13.78	14.06	0.8759	5.479	3.156	3.136	4.872
13.74	14.05	0.8744	5.482	3.114	2.932	4.825
14.59	14.28	0.8993	5.351	3.333	4.185	4.781
13.99	13.83	0.9183	5.119	3.383	5.234	4.781
15.69	14.75	0.9058	5.527	3.514	1.599	5.046
14.7	14.21	0.9153	5.205	3.466	1.767	4.649
12.72	13.57	0.8686	5.226	3.049	4.102	4.914
14.16	14.4	0.8584	5.658	3.129	3.072	5.176
14.11	14.26	0.8722	5.52	3.168	2.688	5.219
15.88	14.9	0.8988	5.618	3.507	0.7651	5.091
12.08	13.23	0.8664	5.099	2.936	1.415	4.961
15.01	14.76	0.8657	5.789	3.245	1.791	5.001
16.19	15.16	0.8849	5.833	3.421	0.903	5.307
13.02	13.76	0.8641	5.395	3.026	3.373	4.825
12.74	13.67	0.8564	5.395	2.956	2.504	4.869
14.11	14.18	0.882	5.541	3.221	2.754	5.038
13.45	14.02	0.8604	5.516	3.065	3.531	5.097
13.16	13.82	0.8662	5.454	2.975	0.8551	5.056
15.49	14.94	0.8724	5.757	3.371	3.412	5.228
14.09	14.41	0.8529	5.717	3.186	3.92	5.299
13.94	14.17	0.8728	5.585	3.15	2.124	5.012
15.05	14.68	0.8779	5.712	3.328	2.129	5.36
16.12	15	0.9	5.709	3.485	2.27	5.443
16.2	15.27	0.8734	5.826	3.464	2.823	5.527
17.08	15.38	0.9079	5.832	3.683	2.956	5.484
14.8	14.52	0.8823	5.656	3.288	3.112	5.309
14.28	14.17	0.8944	5.397	3.298	6.685	5.001
13.54	13.85	0.8871	5.348	3.156	2.587	5.178
13.5	13.85	0.8852	5.351	3.158	2.249	5.176
13.16	13.55	0.9009	5.138	3.201	2.461	4.783
15.5	14.86	0.882	5.877	3.396	4.711	5.528
15.11	14.54	0.8986	5.579	3.462	3.128	5.18
13.8	14.04	0.8794	5.376	3.155	1.56	4.961
15.36	14.76	0.8861	5.701	3.393	1.367	5.132
14.99	14.56	0.8883	5.57	3.377	2.958	5.175
14.79	14.52	0.8819	5.545	3.291	2.704	5.111
14.86	14.67	0.8676	5.678	3.258	2.129	5.351
14.43	14.4	0.8751	5.585	3.272	3.975	5.144
15.78	14.91	0.8923	5.674	3.434	5.593	5.136
14.49	14.61	0.8538	5.715	3.113	4.116	5.396
14.33	14.28	0.8831	5.504	3.199	3.328	5.224
14.52	14.6	0.8557	5.741	3.113	1.481	5.487
15.03	14.77	0.8658	5.702	3.212	1.933	5.439

14.46	14.35	0.8818	5.388	3.377	2.802	5.044
14.92	14.43	0.9006	5.384	3.412	1.142	5.088
15.38	14.77	0.8857	5.662	3.419	1.999	5.222
12.11	13.47	0.8392	5.159	3.032	1.502	4.519
11.42	12.86	0.8683	5.008	2.85	2.7	4.607
11.23	12.63	0.884	4.902	2.879	2.269	4.703
12.36	13.19	0.8923	5.076	3.042	3.22	4.605
13.22	13.84	0.868	5.395	3.07	4.157	5.088
12.78	13.57	0.8716	5.262	3.026	1.176	4.782
12.88	13.5	0.8879	5.139	3.119	2.352	4.607
14.34	14.37	0.8726	5.63	3.19	1.313	5.15
14.01	14.29	0.8625	5.609	3.158	2.217	5.132
14.37	14.39	0.8726	5.569	3.153	1.464	5.3
13.07	13.92	0.848	5.472	2.994	5.304	5.395
13.32	13.94	0.8613	5.541	3.073	7.035	5.44
13.34	13.95	0.862	5.389	3.074	5.995	5.307
12.22	13.32	0.8652	5.224	2.967	5.469	5.221
11.82	13.4	0.8274	5.314	2.777	4.471	5.178
11.21	13.13	0.8167	5.279	2.687	6.169	5.275
11.43	13.13	0.8335	5.176	2.719	2.221	5.132
12.49	13.46	0.8658	5.267	2.967	4.421	5.002
12.7	13.71	0.8491	5.386	2.911	3.26	5.316
10.79	12.93	0.8107	5.317	2.648	5.462	5.194
11.83	13.23	0.8496	5.263	2.84	5.195	5.307
12.01	13.52	0.8249	5.405	2.776	6.992	5.27
12.26	13.6	0.8333	5.408	2.833	4.756	5.36
11.18	13.04	0.8266	5.22	2.693	3.332	5.001
11.36	13.05	0.8382	5.175	2.755	4.048	5.263
11.19	13.05	0.8253	5.25	2.675	5.813	5.219
11.34	12.87	0.8596	5.053	2.849	3.347	5.003
12.13	13.73	0.8081	5.394	2.745	4.825	5.22
11.75	13.52	0.8082	5.444	2.678	4.378	5.31
11.49	13.22	0.8263	5.304	2.695	5.388	5.31

3 Sample Data for dataset 3 (Breast cancer Wisconsin (Diagnostic) data):

att1	att2	att3	att4	att5	att6	att7	att8	……	att30
13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	…….	0.05766
13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	…….	0.06811