NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA

Improving Influenced Outlierness(INFLO) Outlier Detection Method

by

Shashwat Suman

under the guidance of

Prof. Bidyut Ku. Patra

A thesis submitted in partial fulfillment for the
degree of Bachelor of Technology
in the

Department of Computer Science and Engineering

May 2013

NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA

Certificate

This is to certify that the thesis entitled, 'Improving Influenced Outlierness(INFLO) Outlier Detection Method' by Shashwat Suman in partial fulfilment of the requirements for the award of Bachelor of Technology Degree in Computer Science and Engineering at the National Institute of Technology, Rourkela is an authentic work carried out by him under my supervision and guidance. To the best of my knowledge the matter embodied in the thesis has not been submitted to any other University/ Institute for the award of any Degree or Diploma.

Date: Prof Bidyut Ku. Patra

Abstract

Anomaly detection refers to the process of finding outlying records from a given dataset. This process is a subject of increasing interest among analysts. Anomaly detection is a subject of interest in various knowledge domains.As the size of data is doubling every three years there is a need to detect anomalies in large datasets as fast as possible.Another need is the availability of unsupervised methods for the same.
This thesis aims at implement and comparing few of the state of art unsupervised outlier detection methods and propose a way to better them. This thesis goes in depth about the implementation and analysis of outlier detection algorithms such as Local Outlier Factor(LOF),Connectivity-Based Outlier Factor(COF),Local Distance-Based Outlier Factor and Influenced Outlierness. The concepts of these methods are then combined to propose a new method which better the previous mentioned ones in terms of speed and accuracy.

Keywords: Outlier, Anomaly Detection, Data Mining.

Acknowledgements

This project would not have been possible without the help and support of many. I would like to express my gratitude to Prof. Bidyut Patra for his advice during our project work. As my supervisor, he has constantly encouraged me to keep on focused on achieving goal.His vast knowledge and expertise in the area of networking was immensely helpful. His observations and comments helped me to establish to overall direction of the research and to move forward with the study in depth. He has helped us greatly and been a source of knowledge.

I am thankful to all our teachers and friends. My sincere thanks to everyone who has provided us with inspirational words, a welcome ear, new ideas, constructive criticism, and their invaluable time, I am are truly indebted. I must acknowledge the academic resources that we have acquired from NIT Rourkela. I would like to thank the administrative and technical staff members of the department who have been kind enough to advise and help in their respective roles.

Shashwat Suman
(109CS0195)
Department of Computer Science and Engineering
National Institute of Technology
Rourkela

Contents

Certificate i

Abstract ii

Acknowledgements iii

List of Figures v

1 Introduction 1
- 1.1 Applications . . . 1
- 1.2 Defining an Outlier . . . 2
- 1.3 Types of Outliers . . . 3
- 1.4 Output of Outlier Detection . . . 3
- 1.5 Anomaly detection using distance to kth Nearest Neighbor . . . 4
- 1.6 Relative Density . . . 5
- 1.7 Global versus local approaches to outlier detection . . . 6
- 1.8 An analysis of Nearest Neighbor Based Techniques . . . 7

2 Literature Survey 9
- 2.1 Local Outlier Factor(LOF) . . . 9
  - 2.1.1 Definitions . . . 9
  - 2.1.2 Properties of LOF . . . 10
- 2.2 Connectivity Based Outlier Factor(COF) . . . 12
  - 2.2.1 Definitions . . . 12
- 2.3 Local Distance-Based Outlier Detection Factor(LDOF) . . . 15
  - 2.3.1 Formal Definitions . . . 15
- 2.4 Influenced Outlierness(INFLO) . . . 18
- 2.5 A Depth Based Outlier Detection Method . . . 19

3 Objective 21

4 A Proposed Outlier Detection Method 22
- 4.1 Features . . . 22
- 4.2 Pseudocode . . . 22

5 Implementation and Analysis **23**
5.1 Datasets used . . . 23
5.1.1 Dummy Dataset . . . 23
5.1.2 IRIS dataset . . . 24
5.1.3 Spambase dataset . . . 24
5.1.4 Breast Cancer Dataset . . . 25
5.1.5 Seeds Dataset . . . 25
5.2 Results . . . 26
5.2.1 Results on Seeds Dataset . . . 26
5.2.2 Results on Iris 53 Dataset . . . 29
5.2.3 Results on Iris 106 Dataset . . . 32
5.2.4 Breast Cancer Dataset . . . 33
5.2.5 Spambase Dataset . . . 34
5.3 Analysis w.r.t. time . . . 35

6 Conclusion **36**

List of Figures

1.1 . . . 2
1.2 Test Case 1 . . . 5

2.1 A Gaussian Distribution . . . 11
2.2 Values of LOF with varying k . . . 11
2.3 Data Instances conforming to patterns . . . 12
2.4 Nearest Neighborhood in COF . . . 13
2.5 A test case to demonstrate effictiveness of COF . . . 14
2.6 Test case to demonstrate effectiveness of LDOF . . . 15
2.7 Showcasing $d_{x_p}D_{x_p}$ . . . 16
2.8 A data instance is located between two clusters . . . 17
2.9 . . . 18
2.10 Convex Hull on a Gaussian Distribution . . . 19
2.11 Model . . . 20

5.1 LOF Output on Seeds dataset(The last three instances have higher values and thus classified as outliers.) . . . 26
5.2 LDOF Output on Seeds Dataset . . . 26
5.3 INFLO Output on Seeds Dataset . . . 27
5.4 PODM output on Seeds Dataset . . . 27
5.5 Results in Graphical Form . . . 28
5.6 Seeds dataset Table . . . 28
5.7 Results of LOF on Iris53 Dataset . . . 29
5.8 Results of LDOF on Iris53 Dataset . . . 29
5.9 Results of INFLO on Iris53 Dataset . . . 30
5.10 Results of PODM on Iris53 Dataset . . . 30
5.11 Results in graphical form . . . 31
5.12 Iris53 dataset Table . . . 31
5.13 Results on IRIS106 Dataset . . . 32
5.14 Iris106 dataset Table . . . 32
5.15 Results on Breast Cancer Dataset in graphical form. . . . 33
5.16 Breast Cancer dataset Table . . . 33
5.17 Results on Spambase Dataset in graphical form . . . 34
5.18 Spambase dataset Table . . . 34
5.19 Analysis with respect to Time . . . 35

Chapter 1

Introduction

Outlier detection in datasets has been an object of interest throughout history. Outliers can sometimes heavily distort the statistical data and sometimes the effect is hardly noticeable. Some outliers can bring to attention a very important fact e.g. the discovery of Argon resulted from unexpected differences in weight of Nitrogen.

Another example where outlier detection shows its vast importance is security, especially air safety. One of the high jacked airplanes of 9/11 had a particular anomaly. It has five passengers which (i) Were non-US citizens (ii) Had links to a particular country (iii) Had purchased a one-way ticket (iv) Had paid in cash (v) Had no luggage. Of course one or two such passengers in an airplane is a normal observation but 5 in a particular plane is an anomaly. The task of outlier detection is often a safety critical task e.g. aircraft engine rotation. An outlier can be an anomalous object in an image like a landmine or a abandoned briefcase in security tapes. Outlier detection can also spot performance degradation in machines of factories. This helps to identify faults in machines and avoid machine failure or disaster.

1.1 Applications

A more succinct list of outlier detection applications are given below

Fraud detection
Refers to criminal activities involving banks, credit card companies, insurance companies, cell phone companies, stock market etc. Fraud occurs when resources of a company are utilized in an unauthorized way. Companies want to detect these fraudulent transactions before they incur heavy losses.

Insider trading

Insider trading is also an area where outlier detection can be applied. Insider trading refers to people making illegal profits using insider information of companies before it is made public.

Medical health anomaly detection

Outlier detection can be applied to patient records which contains the results of various tests performed on the patient. This can lead to discovery of complications of the patient. Outlier detection also helps in detecting epidemics

Fault diagnosis

Faults in machinery like motors, generators, transformers etc. or instruments in space shuttles. Structural faults like cracked beams or unstable foundations.

Image Processing

Usually in large sized things like images, a lot of unexpected outliers tend to creep in. Anomalous data is very different from normal data and can be removed though outlier detection to give a clear view of the image and its components.

1.2 Defining an Outlier

Definition of Hawkins [Hawkins 1980]: An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism Abnormal objects deviate from this generating mechanism.

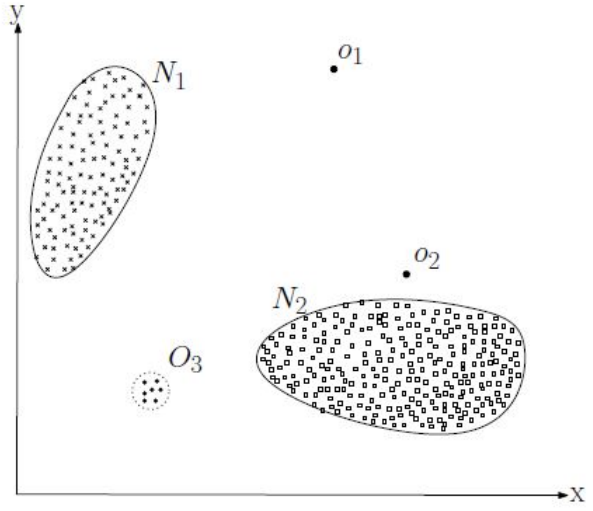


FIGURE 1.1

1.3 Types of Outliers

Outliers often arise due to human carelessness, faults in systems, natural deviation in dataset, fraud etc. However it is important to differentiate between the applications of outlier detection e.g. if there is a clerical error in the data of a , the entry clerk should be notified and the data should be corrected. However in industry when there are faults in a machine and the machine is damaged or in a safety critical system like an intrusion monitoring system or a fraud detection system, an alarm must be sounded to notify the system administrators about the problem. There are three approaches to the problem of outlier detection.

Type 1

In this type outliers are determined with no prior knowledge of dataset. This is a learning process which treats the data as a static distribution which finds remote points and flags them as potential outliers. The main cluster might be subdivided to improve outlier detection. It also assumes that normal instances of data are much more frequent than anomalous instances. A drawback of this methodology is that is needs data to be dynamic and thus needs a very large database. We will mainly be concerned with this type in this thesis.

Type 2

In this type it is assumed that the dataset only has labelled instances of 'normal' class. They do not require labels for anomalous class e.g. in case of safely critical systems where an anomaly would mean an accident which would be hard to model. Typically a model is built for normal behavior and then is used to identify anomalies.

Type 3

In this type there is availability of labeled instances of both normal and anomalous class. In this type when a new instance of data is encountered it is compared to the model to determine its class. One of the drawbacks of this class is obtaining labels for anomaly classes is hard.

1.4 Output of Outlier Detection

Reporting of anomalies is important aspect, generally outputs are of two types

Labels

In this the data instances are assigned a 'label' which term it as normal or anomalous to each instance. Thus it gives a binary output.

Score
In this data instances are assigned an anomaly score which basically indicates the outlierness of the object or the degree to which the object is an outlier. So the output is basically points with their degree of outlierness. An analyst might analyze the top few outliers or propose a 'cut off' to select the anomalies. Thus it gives a continuous output.

Many anomaly detection techniques prefer using the 'top-n' outlier process so they prefer score. If binary labelled data is required ,a threshold(lower bound) can be applied to the scores to give a binary output.

1.5 Anomaly detection using distance to kth Nearest Neighbor

A very naive anomaly detection technique is the anomaly score of a data instance,it is defined as its distance to its kth nearest neighbor in a given data set. Usually k is chosen to be greater than 1.

Another way to computer the anomaly score is to fix the number of nearest neighbors say n which are at a distance d apart. This is like computing the global density of a data instance in a hypersphere of radius d. In 2 dimensional data density-(n/pie*dsquare) can be used as the density. The inverse of density is then used as the score. In computing this density some methods fix the neighbors and use 1/d as the score and some method fix d and use 1/n as the score.

The problem with using such methods.

Test Case 1

Consider a 2 dimensional dataset as shown in figure 1.2. Let C_1 and C_2 be two major clusters containing 100 and 400 data instances respectively. It can be seen in the figure that C_1 cluster is a lot more sparse than cluster C_2.We have two additional points P_1 and P_2 which we will be focusing on. As we can see both are outliers and have been formed by mechanisms other than normal. If we apply by k nearest neighbor technique to P_1 we will classify it as an outlier as its n nearest neighbors are very far. But in the case of P_2 its n nearest neighbors are not that far and are as sparse as data instance in

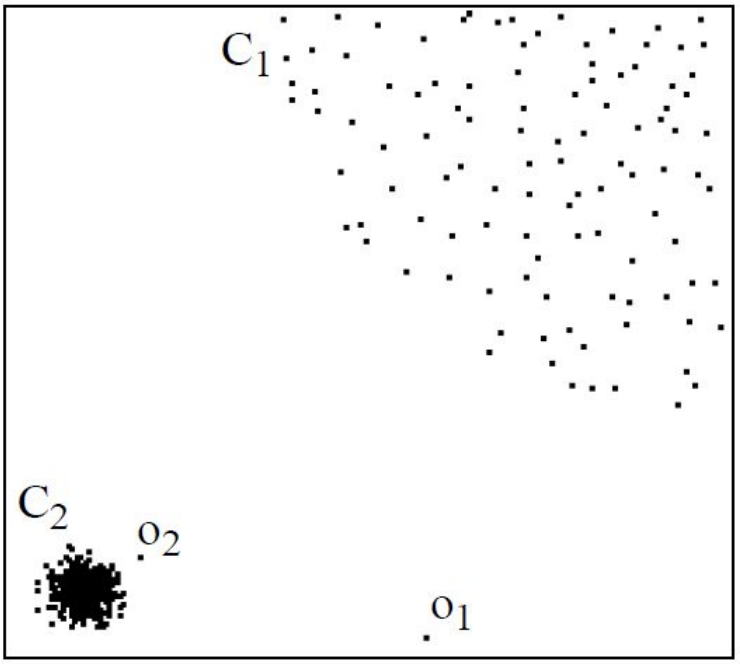


FIGURE 1.2: Test Case 1

C_1.So by using the simple k nearest neighbor method we will wrongly classify it as a normal instance when it is not.

E.g. $DB(\epsilon, \pi)$-outlier model Parameters ϵ and π cannot be chosen so that O_2 is classified as an outlier but none of the points in cluster C_1 are classified as outliers.

In detection mechanism which are just based on KNN-distance like-
Taking the kNN distance of a point as its outlier score [Ramaswamy et al 2000]

Aggregating the distances of a point to all its 1NN, 2NN, , kNN as an outlier score [Angiulli and Pizzuti 2002]

So we conclude that in the dataset of varying density we need a better method to detect outliers. Thus we bring the concept of using relative density.

1.6 Relative Density

As we saw in the previous, section simply using the distance to the kth nearest neighbor is not enough to classify an instance especially in the case of varying data instances. We need to take in account the local and relative density of the point rather than just the global density.

To solve this problem a new anomaly score was given to a data instance which is called

LOF(Local Outlier Factor).LOF takes into account both the density of the given data instance and the density of the data instances in the K nearest neighbor set of point. It is the ratio of the average local density of the k nearest neighbors to the point itself. To find this first the k nearest neighbors of the point is computed. The local density is computer by diving k by the radius of the hypersphere. For a data instance lying in a dense neighborhood its density will be similar to its k nearest neighbor whereas an isolated point will have a high density compared to its k nearest neighbors and will have a high LOF score. In figure 1.1 LOF will correctly classify both P1 and P2 as outliers .However LOF is not this simple to compute as we will see in further sections where we will study it in more detail.

Later another anomaly score technique was proposed call COF(Connectivity-Based Outlier Factor).The basic difference between COF and LOF only lies in the procedure for calculation of the k nearest neighbor set. In COF the k nearest neighbor set is calculated incrementally. First, the closest instance is added to the k nearest neighbor set.Then the next point closest to the set is added to the set. This is continued till the set contains k instances which contains the k neighborhood of the point. COF is useful is capturing shapes and patterns like lines and captures it better than LOF.

Other methods like ODIN(Outlier Detection using in-degree number) is used for each dataset. For a given data instance ODIN is the number of points who contain the original point in their k nearest neighbor set. The inverse of ODIN is generally used as the anomaly score. This is discussed in detail under the INFLO outlier detection technique. Other methods include MDEF(Multi-granularity Deviation Factor) which for a given data instance's MDEF is equal to the standard deviation in the local densities of the local neighborhood density. The inverse of this standard deviation is taken as the anomaly score. Other detection techniques like LOCI finds anomalous micro clusters.

1.7 Global versus local approaches to outlier detection

This means the type of reference set we are ready to take with respect to the outlierness of an object.

Global approach

The basic premise of this approach is that the reference set contains all other data object i.e. there is one dataset with all the objects in the reference set. Here we are assuming that there is only one mechanism of generation of datasets. However this

might not always be the case. Another drawback is that there might be outliers in the reference set which might distort the result.

Local approach

The basic premise of this approach is that the reference set contains a small subset of dataset. We are not assuming a single mechanism here but there is no defined method to choose a proper reference set.

Some approaches lie somewhere between global and local approach.

Types of Anomaly Detection Techniques

1. Statistical Tests
2. Depth-based Approaches
3. Deviation-based Approaches
4. Distance statistical model
5. Distance-based Approaches
6. Density-based Approaches

1.8 An analysis of Nearest Neighbor Based Techniques

The advantages of nearest neighbor based techniques are as follows

1. The main advantage of this is that it is unsupervised and can run only with the given data.
2. Different kinds of data can be used, only the equation for finding the distance has to be changed. Different number of attributes can be adjusted accordingly. endenumerate

   The disadvantages of nearest neighbor based techniques are as follows:

   (a) If there are a lot of anomalies in the dataset there might be misclassification of data and anomalous data might be classified as normal. i.e. false positive rate will be very high.

   (b) The computational complexity of such method is high as the neighborhood sets of all the data instances have to be calculated.

(c) Defining the method of calculating distances is also a challenge. In regular datasets Euclidean distance is preferred but when the data is complex like graphs or sequences defining distances is a problem.

Chapter 2

Literature Survey

2.1 Local Outlier Factor(LOF)

We described in fig 1.2 the problem with simplistic use of the k nearest neighbor procedure leads to false labelling. Local density based methods compare the local density of the object to that of its neighbors. For the LOF to accomplish that the following definitions were used.

2.1.1 Definitions

K-Dist of an object p

For any positive integer k, the k-distance of object p, denoted as k-dist(p), is defined as the distance d(p,o) between p and an object o such that $o \epsilon D$ is

(i) for at least k objects $ó \epsilon D / \{p\}$ it holds that $d(p, ó) \leq d(p, o)$, and
(ii) for at most k-1 objects $ó \epsilon D / \{p\}$ it holds that $d(p, ó) < d(p, o)$.

K-distance neighborhood of an object p

Given the k-distance of p, the k-distance neighborhood of p contains every object whose distance from p is not greater than the k-distance, i.e.

$N_{k-distance(p)}(p) = \{q \epsilon D \setminus \{p\} \mid d(p, q) \leq k - distance(p)\}$

These objects q are called the k-nearest neighbors of p.

Reachability distance of an object p w.r.t. object o

Let k be a natural number.
$reach - dist_k(p,o) = max\{k - distance(o), d(p,o)\}$

Local Reachability Density of an object p

$$lrd_{Minpts}(p) = 1/\left\{\frac{\sum_{o\epsilon N_{Minpts}(p)} reach{-}dist_{Minpts}(p,o)}{N_{Minpts}(p)}\right\}$$

Local outlier factor of an object p

$$LOF_{Minpts}(p) = \frac{\sum_{o\epsilon N_{Minpts(p)}} \frac{lrd_{Minpts}(o)}{lrd_{Minpts}(p)}}{|N_{Minpts}(p)|}$$

2.1.2 Properties of LOF

- LOF $\simeq$ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
- LOF $\gg$ 1: point is an outlier.
- The output factor depends a lot on the choice of k(Minpts).

Here there is a single parameter Minpts which can vary the LOF.Let us see its impact on LOF- Let take a Gaussian distribution as shown in fig 2.1

The following figure shows the distribution of values of LOF with varying k(2-50).As we can see the standard deviation of LOF values only really stablises when $k > 10$.

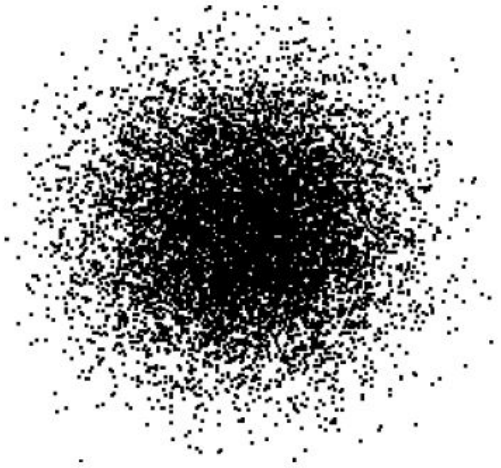


FIGURE 2.1: A Gaussian Distribution

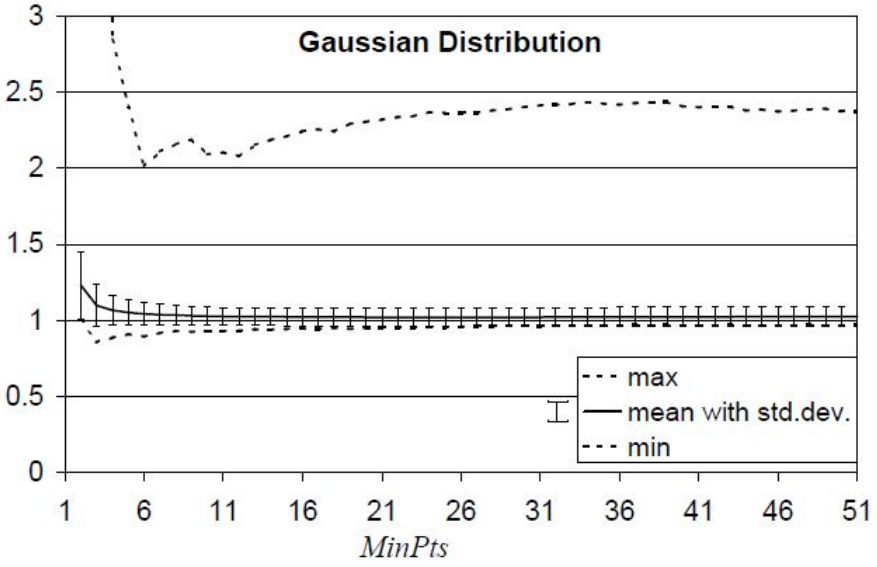


FIGURE 2.2: Values of LOF with varying k

2.2 Connectivity Based Outlier Factor(COF)

The connectivity-based outlier factor is a local density based approach proposed in order to find outliers when data specifies certain patterns like lines or spheres. It is a simple variant of LOF which also uses the k nearest neighbor method but method to calculate k nearest neighbor is different. So basically COF can detect outliers in low density patterns and thus is more effective in such a case than LOF. An example of such patterns are straight lines.

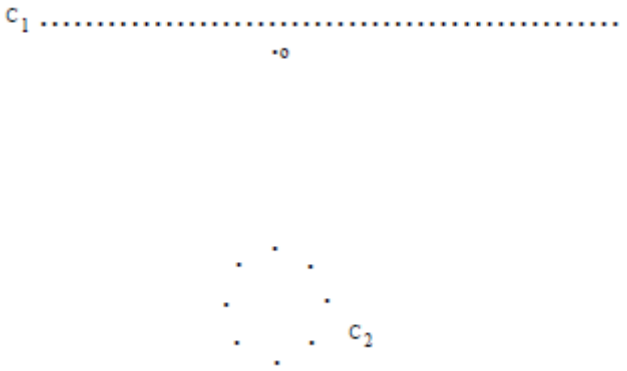


FIGURE 2.3: Data Instances conforming to patterns

In the fig 2.3 it can be noticed that the points in the spherical cluster are far apart as compared to the straight line cluster. As we see the Anomaly point here is point O which belongs to neither of the clusters. But as we can clearly see the k near neighborhood of o will have many points and thus will have an LOF which is close to any point in the spherical cluster, thus it gets wrongly classified as a normal data instance. Thus COF was introduced to find this pattern.

The local density in the COF is the inverse of the average chaining distance. The average chaining distance is different from the local reachability distance in LOF as it takes into account the distance of each point from the original point and the sequence. The following figure helps in explaining the chaining distance method to calculate the k nearest neighbor set of p .

2.2.1 Definitions

Distance between two sets of points

d(P,Q) The distance between the set P and Q is the minimum distance between their elements, denoted be d(P,Q)

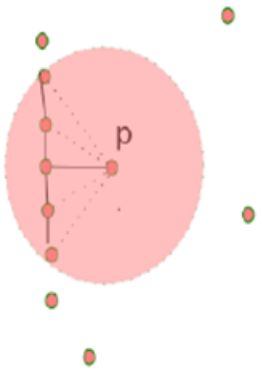


FIGURE 2.4: Nearest Neighborhood in COF

Let P,Q $\subseteq$ D, P $\cap$ Q =ϕ and P,Q $\neq \phi$. For any given q $\in$ Q we say that q is the nearest neighbor of P in Q if dist (q, P) = dist(Q, P)

Set-Based Nearest Path(SBN)

A set-based nearest path, or SBN-path, from p_1 on G is a sequence $(p_1, p_2, ...p_r)$ of all the elements in G such that for all $1 \leq i \leq r-1, p_{i+1}$ is a nearest neighbor of set $\{p_1, ., p_i\}$ in $\{p_{i+1}, ., p_r\}$

Average Chaining Distance

Let $s = (p_1, p_2, ...p_r)$ be an SBN-path from p_1 on G. The average chaining distance from p_1 on G, denoted by $ac - distG(p_1)$, is defined as

$ac - DistG(p_1) = \sum_{i=1}^{r} \frac{2(r-1)}{r(r-1)} dist(e_i)$
where $dist(e_1) = dist(p_1, p_2, ...p_i, (p_{i+1}, ...p_r))$

Connectivity-Based Outlier Factor(COF)

$\text{COF}_k(p) = \frac{N_k(p).ac-dist(p)}{\sum_{o \epsilon N_K(p)} ac-dist(o)}$

As observed by the formulas the average chaining distance is the weighted sum of the cost description sequence. The earlier edges have a greater contribution to the sum than the latter edges. Like LOF a score near to 1 indicates that the point is not an outlier. A score much greater than 1 indicates outlierness.

TEST CASE 2

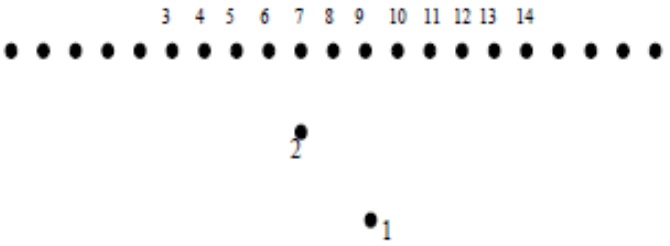


FIGURE 2.5: A test case to demonstrate effictiveness of COF

We take another test case in which the dataset mainly consists of one straight line of data as shown in the figure.There are two outlying points 1 and 2 and we apply the COF algorithm to all points.The following result was obtained with k = 10, we have the following:

COFk(1) = 2.1
COFk(2) = 1.35
COFk(3) = 1.11
COFk(4) = 1.07
COFk(5) = 1.06
COFk(6) = 1.00
COFk(7) = 0.96
COFk(8) = 1.00

Thus we identify the correct outliers using COF method.

2.3 Local Distance-Based Outlier Detection Factor(LDOF)

The previous outlier detection schemes are average when it comes to detecting outliers in real world scattered datasets. LDOF uses the relative distance from an object to its neighbors to measure how much objects deviate from their scattered neighborhood. The higher the factor is the more likely the point is an outlier. It is observed that outlier detection schemes are more reliable when used in a top-n manner. This means that the top n factors are taken as outliers, the n is decided by the user as per his requirements.

TEST CASE 3

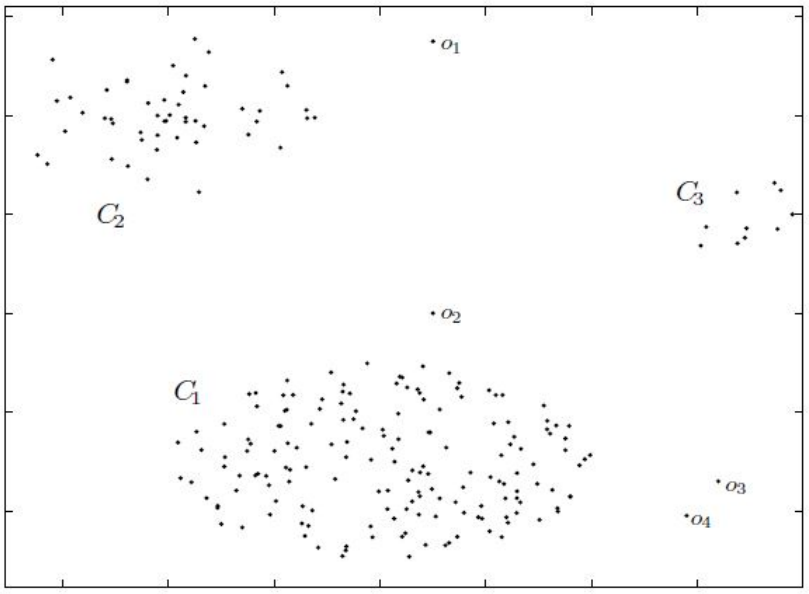


FIGURE 2.6: Test case to demonstrate effectiveness of LDOF

We use a test case in which there are 3 major clusters C_1,C_2 and C_3 and four outlying points O_1,O_2,O_3,O_4.When we set a value ok $k > 10$ the cardinality of a cluster. i.e. in this case C_3 is the smallest cluster whose cardinality is 10.If we set $k > 10$ we get wrong values of KNN and LOF as it starts taking points from clusters C_1 and C_2.We solve this proposing the following method.

2.3.1 Formal Definitions

KNN distance of x_p

Let N_p be the set of the k-nearest neighbors of object x_p (excluding x_p). The k-nearest neighbors distance of x_p equals the average distance from x_p to all objects in N_p. More formally, let $dist(x, x) > 0$ be a distance measure between objects x and ó. The k-nearest

neighbors distance of object x_p is defined as

$\bar{d}_{xp} = \frac{1}{k} \sum_{x_i \epsilon N_p} dist(x_i, x_p)$

KNN inner distance of x_p

Given the k-nearest neighbors set N_p of object x_p, the k-nearest neighbors inner distance of x_p is defined as the average distance among objects in N_p

$\bar{D}_{xp} = \frac{1}{k(k-1)} \sum_{x_i, \acute{x}_i \epsilon N_p, i \neq \acute{i}} dist(x_i, x_{\acute{i}})$

LDOF of x_p

$LDOF(x_p) = \frac{\bar{d}_{xp}}{\bar{D}_{xp}}$

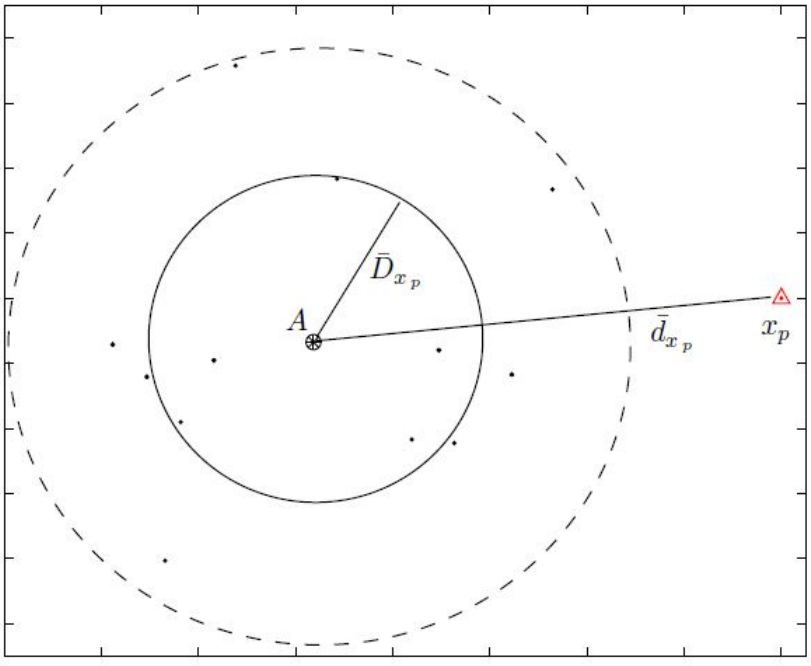


FIGURE 2.7: Showcasing $d_{x_p} D_{x_p}$

Here d mainly denotes the distance between the specified point and all other points in the k neighborhood set of the point and D represents the distance between the points in the k neighborhood set. Another way to minimize calculation is to find the median of the k neighborhood points and name is x and then calculate distance between points in k neighborhood set and $\acute{x}$. This minimizes calculations.

Furthermore LDOF is often used as top-n LDOF and it is used as the following-

Input: A given dataset D, natural numbers n and k.

(a) For each object p in D, retrieve p's k-nearest neighbors

(b) Calculate the LDOF for each object p.The objects with $LDOF < LDOF_{lb}$ are directly discarded

(c) Sort the objects according to their LDOF values

(d) Output: the first n objects with the highest LDOF values

Complexity of the LDOF algorithm only relies on the computation of k nearest neighbors. It is naively done in $O(n^2)$, however using data structures like X-tree or R-tree the complexity can be reduced to O(n2).Later the LDOF values can be sorted by merge sort to find the top-n values.

A problem with LDOF

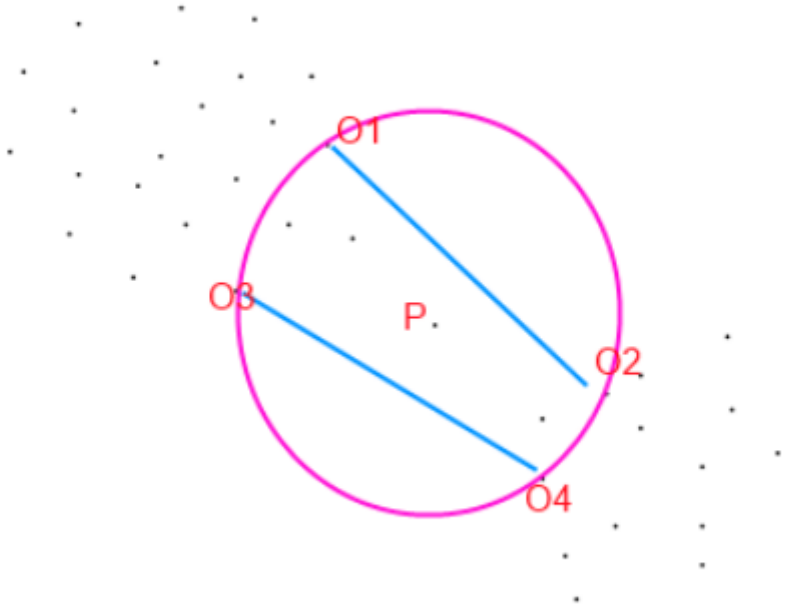


FIGURE 2.8: A data instance is located between two clusters

When a data instance is located between two clusters.The denominator value D increases abnormally as the interdistance between the objects of the K nearest neighborhood increases.This leads to a low factor and leads false classification of the instance as an outlier. Thus LDOF has a high False Positive Rate

2.4 Influenced Outlierness(INFLO)

INFLO was introduced in 2006. It is also based on LOF, however it expands the neighbor- hood of the object to the inuence space (IS) of the object. INFLO was introduced in order to handle the case where clusters with varying densities are in close proximity. Figure 2.9 shows an example of such a case. The data sets has two clusters C_1 and C_2, where C_1 is more dense than C_2. Point p for instance would have the same or an even higher LOF score when k is equal to 3 as point q. This is because the nearest neighbors of p all lie within cluster C_1 as shown in the figure. This is counter intuitive as point p actually lies within cluster C_2. The inuence space overcomes that problem by taking more neighbors into account, namely the reverse k nearest neighbors set (RNN_k). $RNN_k(p)$ is the set of objects that has p in its k-neighborhood set. This is shown in figure 3.4 where s and t are the reverse neighbors of p.The definitions of RNN_k and IS are given below.

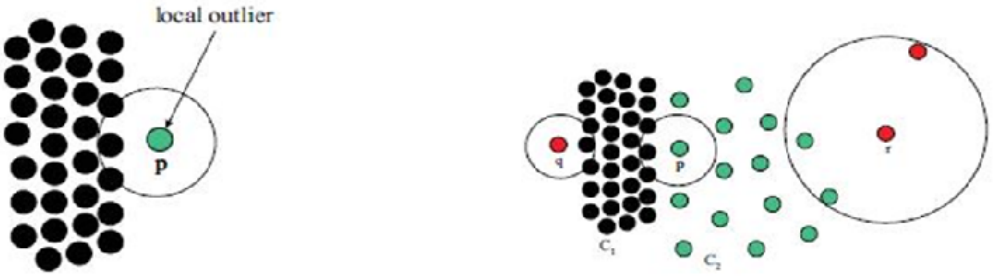


FIGURE 2.9

Formal Definitions

- Reverse k Nearest Neighbor set (RNN)
  $RNN_k(p) = \{q \mid q \epsilon Z, p \epsilon NN_k(p)\}$
- Local density of P
  $den(p) = \frac{1}{K_{dist(p)}}$
- Influence Space (IS)
  $\text{IS}_k(p) = RNN_k(p) \cup NN_k(p)$
- $\text{INFLO}_k(p) = \frac{den_{avg}(IS_k(p))}{den(p)}$
- Where $\text{den}_{avg}(IS_k(p)) = \frac{\sum_{o \epsilon IS_k(p)} den(o)}{IS_k(p)}$

As with other outlier detection method, if $INFLO >> 1$ the point is classified as an outlier.

2.5 A Depth Based Outlier Detection Method

Motivation

Need a method to detect outliers on the fringe portions of dataspace but independent of the distribution of the dataspace.

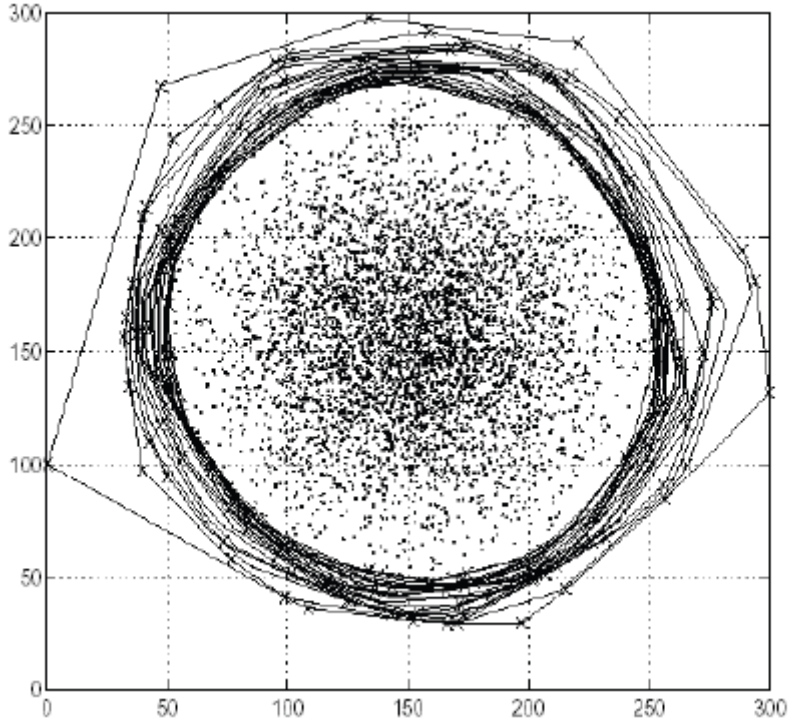


FIGURE 2.10: Convex Hull on a Gaussian Distribution

Basic Idea

Data objects are organized in convex hull layers. Objects on the outermost layers are Outliers and normal objects are located in the centre of dataspace.

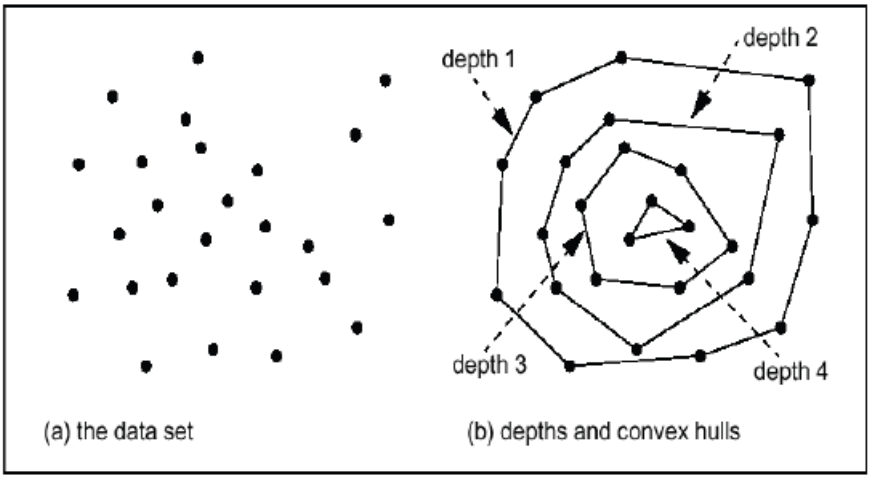


FIGURE 2.11: Model

Model

- Points on the convex hull of the full data space have depth = 1
- Points on the convex hull of the data set after removing all points with depth = 1 have depth = 2 , point on the convex hull of dataset after removing all points having depth = 2 have depth 3 and so on.
- Points having a depth k(as set by the user) are reported as outliers.

Chapter 3

Objective

- Reducing the False Positive error rate as compared to that of LDOF.
- Reducing the False Negative rate than that of LOF.
- Finding a way to improve the Speed of INFLO .
- To improve the efficiency of density based outlier detection and comparison with the existing algorithms

  Let O be the set of outliers

  Let be the set of detected outliers

  Maximize$(O \cap \hat{O})$

  Minimize$(O - \hat{O})$ **(False Positive)**

  Minimize $(\hat{O} - O)$ **(False Negative)**

Chapter 4

A Proposed Outlier Detection Method

4.1 Features

- Uses the concept of d and D ie KNN distance and KNN inner distance of point x_p
- Calculates a temporary factor called TF of the whole dataset and passes top N values(to be set) by the user.
- Also uses the concept of reverse KNN and Influence Space.

4.2 Pseudocode

Algorithm 1

Z=dataset
For each j in Z
Calculate KNN(i)
$X \leftarrow \frac{1}{k} \sum_{x_i \epsilon N_p} dist(x_i, x_p)$
$Y \leftarrow \frac{1}{k(k-1)} \sum_{x_i, \acute{x}_i \epsilon N_p, i \neq \acute{i}} dist(x_i, x_{\acute{i}})$
$LF\,[j] = \frac{X}{Y}$
End For
Sort LF Array
LF-Subset=FIRST(LF,n)
For each a in LF-Subset
Calculate RNNk(a)
$IS_k(a) = RNN_k(a) \cup NN_k(a)$
$PDOM\,[a] = \frac{den_{avg}(IS_k(a))}{den(a)}$

Chapter 5

Implementation and Analysis

Local Outlier Factor(LOF),Connectivity-Based Outlier Factor,Influenced Outlierness(INFLO) and Proposed Outlier Detection Method was implemented under the following specifications.

Language	C++
Processor	Intel(R) Core(TM) i7-2630 CPU @ 2.00GHz
Installed memory (RAM)	4.00 GB
System type:	64-bit Operating System, x64-based processor

5.1 Datasets used

5.1.1 Dummy Dataset

A Dummy test set was taken with very skewed values:

No. of Instances: 9
No. of Attributes: 2

23 56
24 27
65 78
35 45
68 45
57 87
34 76
14 18
99999 99999

The aim of taking this dataset was to check whether the algorithms were working perfectly.

5.1.2 IRIS dataset

Attribute Information

Number of Instances :150
Number of Attributes :4

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

– Iris Setosa
– Iris Versicolour
– Iris Virginica

First 50 instances from Iris Setosa were taken and 3 instances from class Iris Versicolour were taken. The aim of doing so was to test whether the 3 instances from class Iris Versicolour were classified as outliers.

Secondly 50 instances from both Iris Setosa,Iris Versicolour were taken and 6 instances were taken from Iris Virginica.

5.1.3 Spambase dataset

Attribute Information

No. of Attributes: 57
No. of instances: 4601

The last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

5.1.4 Breast Cancer Dataset

Number of Instances: 286
Number of Attributes: 9

Attribute Information

1. Class: no-recurrence-events, recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: yes, no.
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. irradiat: yes, no.

5.1.5 Seeds Dataset

Attribute Information

Number of Instances: 210
Number of Attributes: 7

To construct the data, seven geometric parameters of wheat kernels were measured:

1. area A,
2. perimeter P,
3. compactness C $= 4 * pi * A/P^2$,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

All of these parameters were real-valued continuous.

5.2 Results

5.2.1 Results on Seeds Dataset

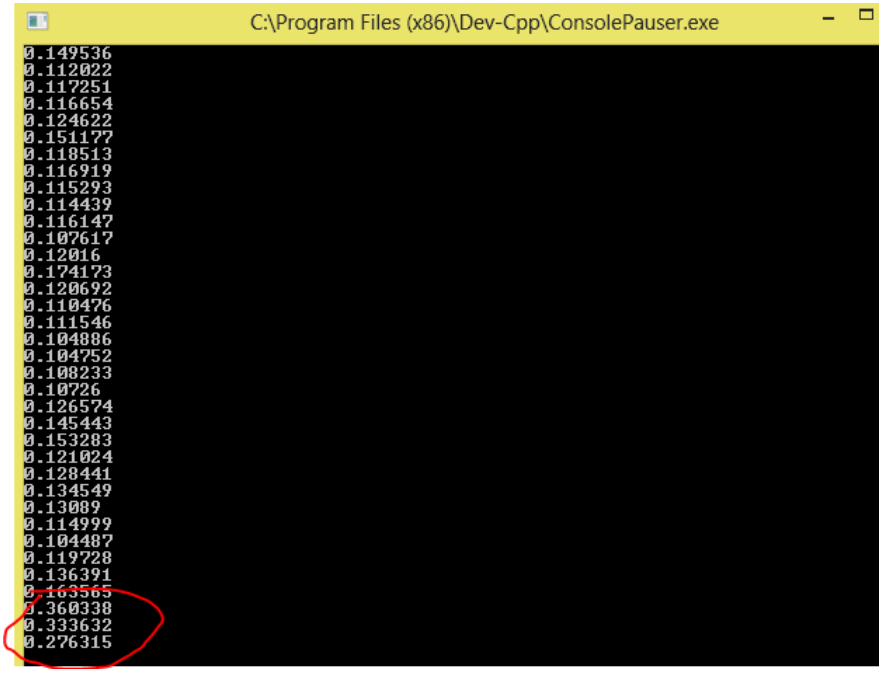


FIGURE 5.1: LOF Output on Seeds dataset(The last three instances have higher values and thus classified as outliers.)

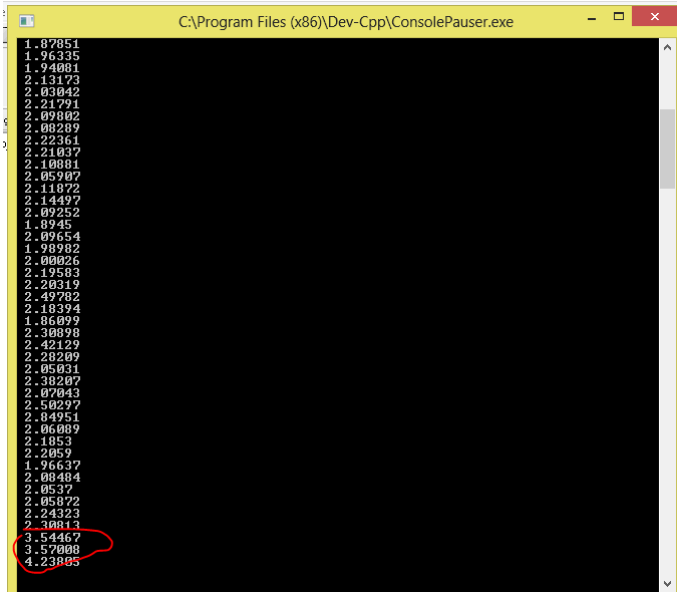


FIGURE 5.2: LDOF Output on Seeds Dataset

FIGURE 5.3: INFLO Output on Seeds Dataset

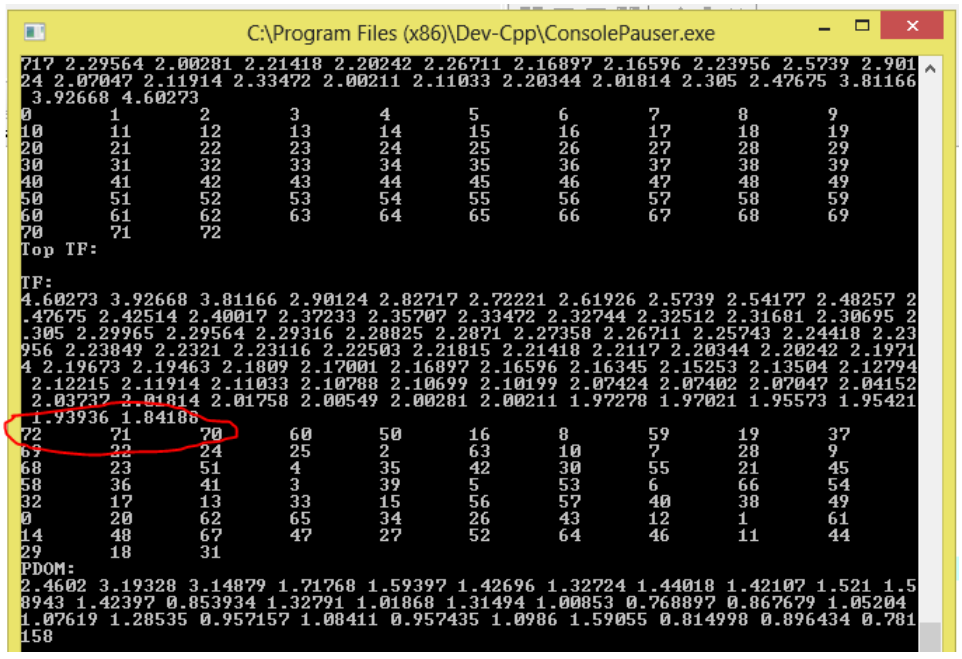


FIGURE 5.4: PODM output on Seeds Dataset

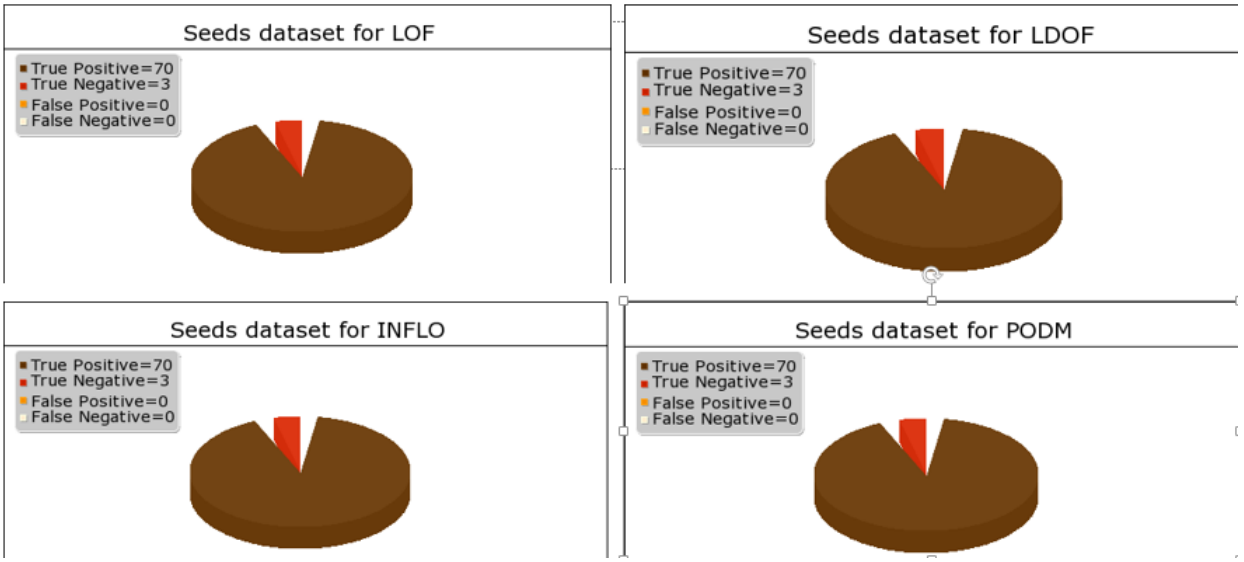


FIGURE 5.5: Results in Graphical Form

	Sensitivity	Specificity
LOF	100%	100%
LDOF	100%	100%
INFLO	100%	100%
PODM	100%	100%

FIGURE 5.6: Seeds dataset Table

5.2.2 Results on Iris 53 Dataset

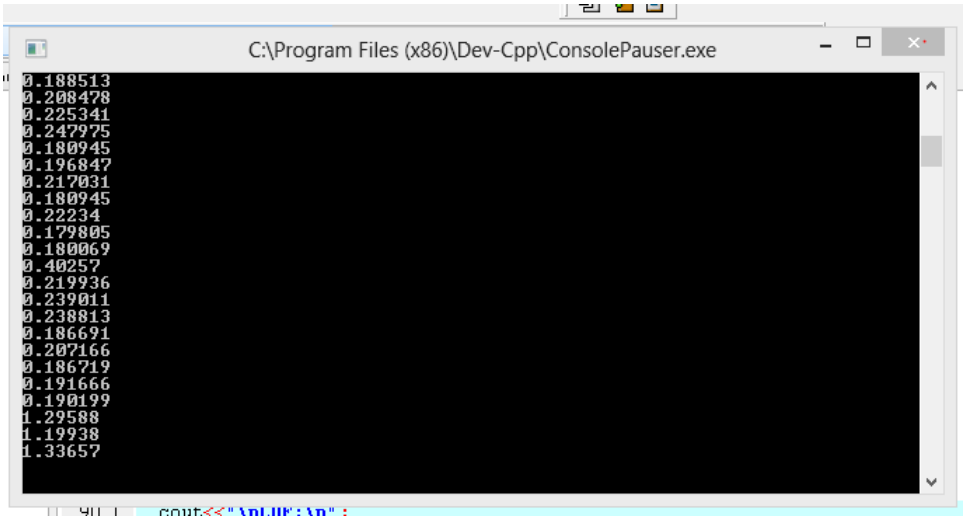


FIGURE 5.7: Results of LOF on Iris53 Dataset

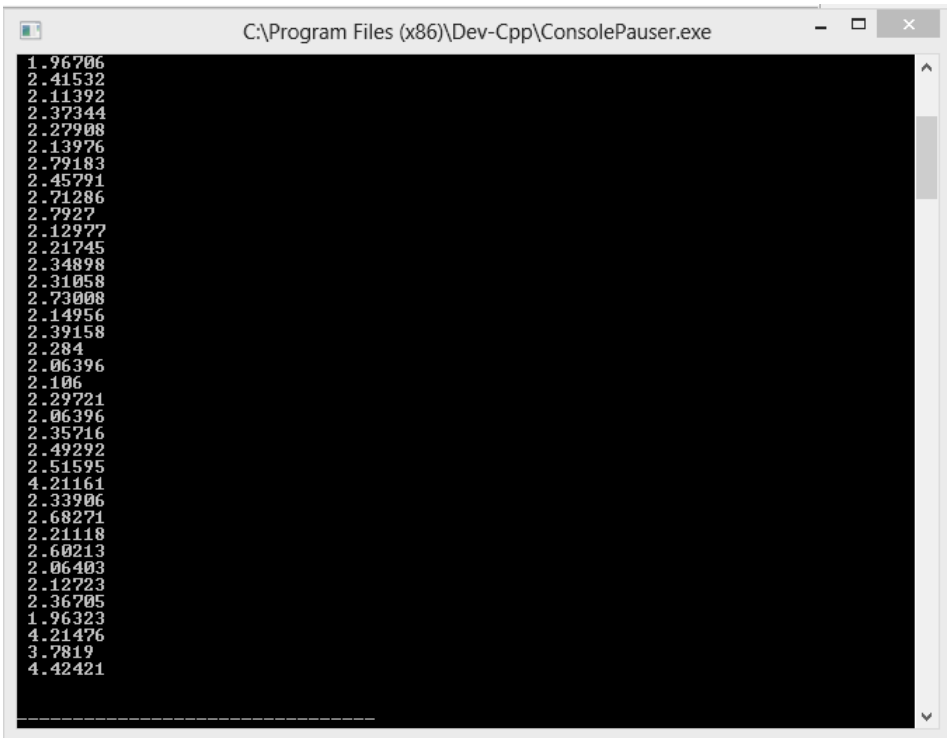


FIGURE 5.8: Results of LDOF on Iris53 Dataset

FIGURE 5.9: Results of INFLO on Iris53 Dataset

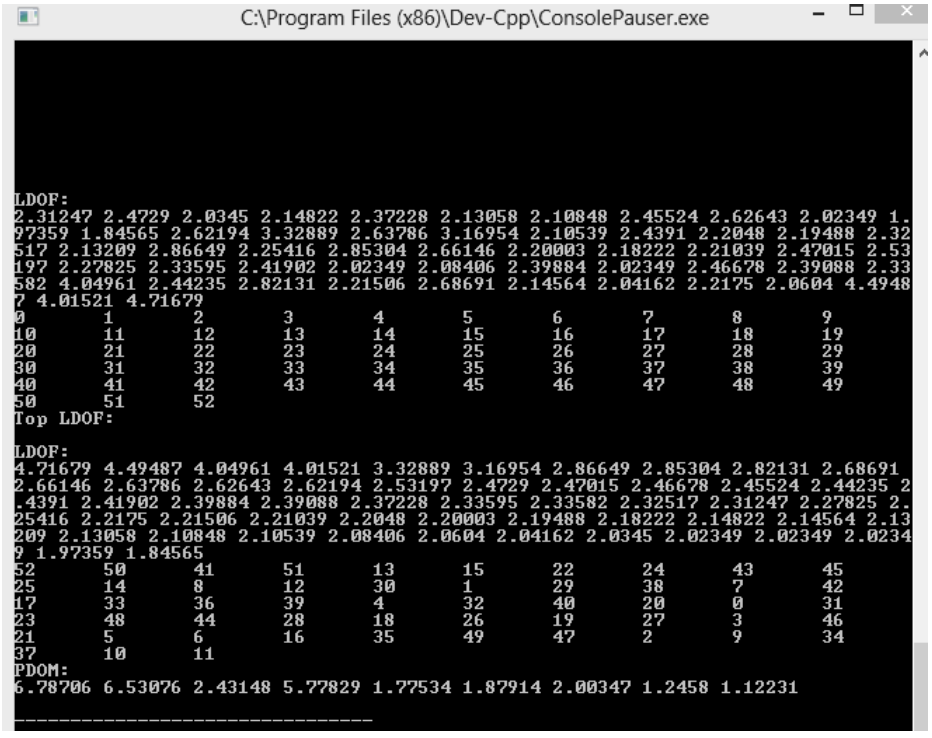


FIGURE 5.10: Results of PODM on Iris53 Dataset

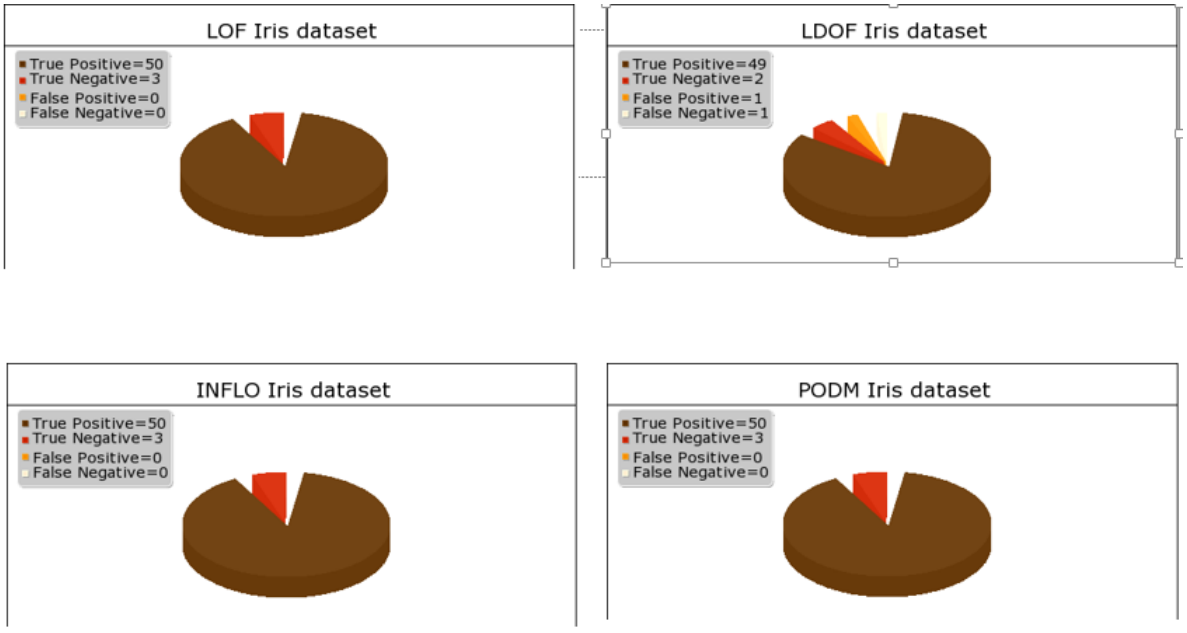
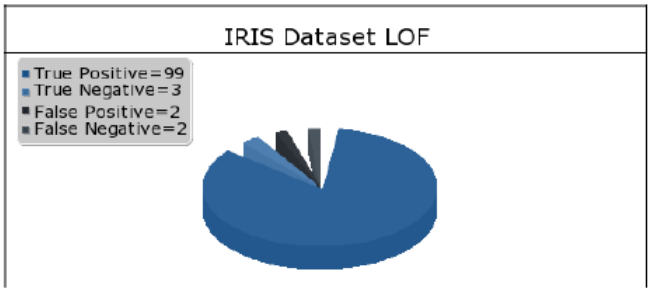


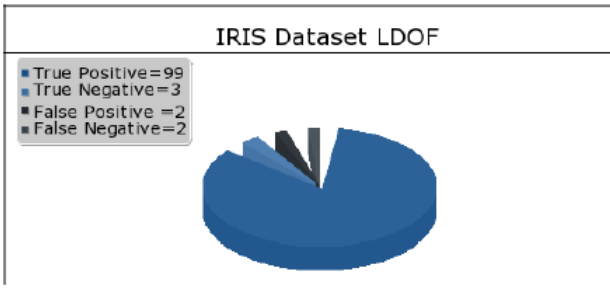
FIGURE 5.11: Results in graphical form

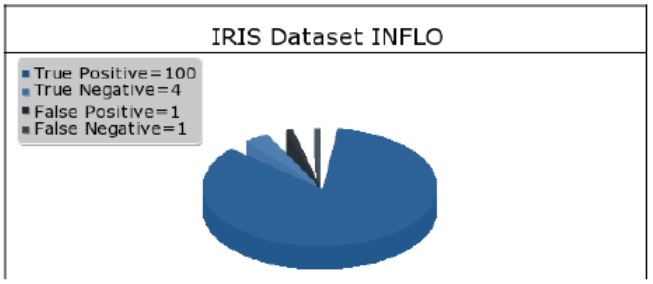
	Sensitivity	Specificity
LOF	100%	100%
LDOF	98%	66%
INFLO	100%	100%
PODM	100%	100%

FIGURE 5.12: Iris53 dataset Table

5.2.3 Results on Iris 106 Dataset







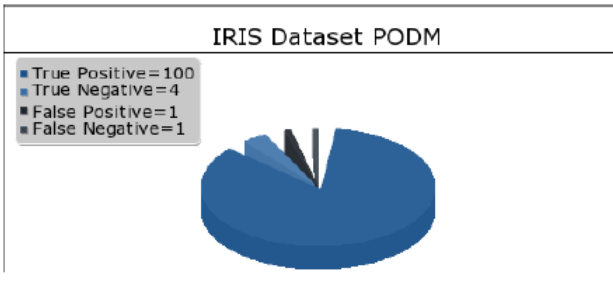


FIGURE 5.13: Results on IRIS106 Dataset

	Sensitivity	Specificity
LOF	98.01%	60%
LDOF	98.01%	60%
INFLO	99.00%	80%
PODM	99.00%	80%

FIGURE 5.14: Iris106 dataset Table

5.2.4 Breast Cancer Dataset

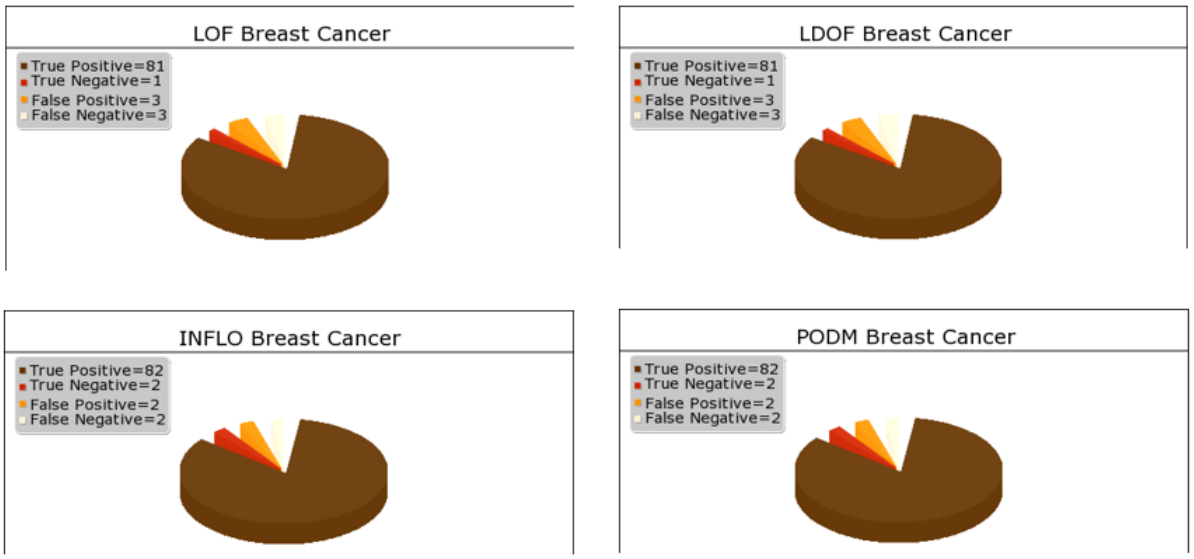


FIGURE 5.15: Results on Breast Cancer Dataset in graphical form.

	Sensitivity	Specificity
LOF	96.42%	25%
LDOF	96.42%	25%
INFLO	97.61%	50%
PODM	97.61%	50%

FIGURE 5.16: Breast Cancer dataset Table

5.2.5 Spambase Dataset

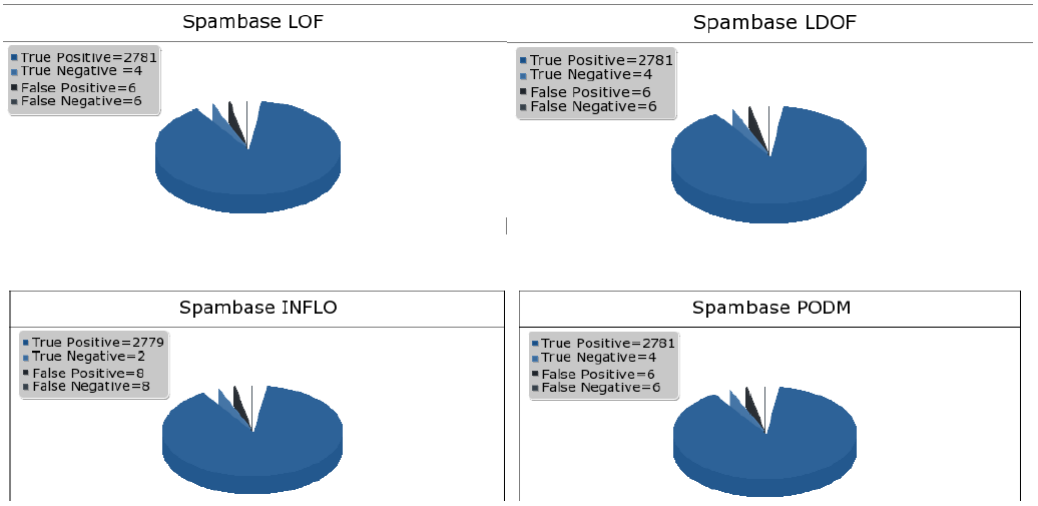


FIGURE 5.17: Results on Spambase Dataset in graphical form

	Sensitivity	Specificity
LOF	99.78%	40%
LDOF	99.78%	40%
INFLO	99.71%	20%
PODM	99.78%	40%

FIGURE 5.18: Spambase dataset Table

5.3 Analysis w.r.t. time

	Iris 53	Iris 106	Breast Cancer 86	Seeds 73	Spambase 2797
LOF	24.00	44.00	62.00	37.00	8096.00
LDOF	.027000	.050000	.047000	.051000	8.55500
INFLO	.043100	.067000	.046000	.143000	9.72000
PODM	.029000	.051000	.043000	.110000	8.74600

FIGURE 5.19: Analysis with respect to Time

Chapter 6

Conclusion

Thus we conclude that the Proposed Outlier Detection Method(PDOM) improves the accuracy of outlier detection wrt LDOF and betters the time taken wrt INFLO thus increasing accuracy and decreasing time taken for execution.This is only achieved by reducing the number of reverse KNN computations.

Bibliography

[1] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. 2000. Lof: identifying density-based local outliers. In Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. ACM Press, 93-104.

[2] Jian Tang, Zhixiang Chen and W.Cheung, D. 2002. Enhancing effectiveness of outlier detections for low density patterns. In Proceedings of the Pacic-Asia Conference on Knowledge Discovery and Data Mining.Pages 535-548.

[3] Ke Zhang and Marcus Hutter and Huidong Jin1;A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data RSISE, Australian National University

[4] Tang, J., Chen, Z., Fu, A. W., and Cheung, D. W. 2006. Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowledge and Information Sys-tems 11, 1, 45-84.

[5] Wen Jin, Anthony K. H. Tung, Jiawei Han, and Wei Wang.Ranking Outliers Using Symmetric Neighborhood Relationship.KDD 2006

[6] VARUN CHANDOLA,ARINDAM BANERJEE and VIPIN KUMAR,Outlier Detection : A Survey,University of Minnesota,ACM Computing 2009