NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA

Study of Distance-Based Outlier Detection Methods

by

Jyoti Ranjan Sethi

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

Under the guidance of
Prof. Bidyut Kumar Patra
Department of Computer Science and Engineering

June 2013

Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, India. www.nitrkl.ac.in

Certificate

This is to certify that the work in the project entitled *Study of Distance-Based Outlier Detection Methods* by Jyoti Ranjan Sethi, bearing roll number 109cs0189, is a record of an original research work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Bachelors of Technology* in *Computer Science and Engineering* during the session of 2012-2013. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Prof. B. K. Patra

Acknowledgements

I express my sincere gratitude to Prof. B. K. Patra and Prof. Korra Sathya Babu for his motivation during the course of the project which served as a spur to keep the work on schedule. I convey my regards to all faculty members of Department of Computer Science and Engineering, NIT Rourkela for their valuable guidance and advices at appropriate times. Finally, I would like to thank my friends for their help and assistance all through this project.

Jyoti Ranjan Sethi

Abstract

An Outlier is an observation which is different from the others in a sample. Usually an anomaly occurs in every data due to measurement error. Anomaly detection is identifying anomalous data for given dataset that does not show normal behavior. Anomaly detection can be classified into three categories: Unsupervised, Supervised and Semi-supervised anomaly detection. Anomaly detection is used variety of domains like fault detection, fraud detection, health monitoring system, intrusion detection. The outlier detection can be grouped into 5 main categories: statistical-based approaches, depth-based approaches, clustering approaches, distance-based approaches and density-based approaches. Distance -based methods i.e. Index-based algorithm, Nested-loop algorithm and LDOF are discussed. To reduce the false positive error in LDOF, we proposed MLDOF algorithm. We tested LDOF and MLDOF by implementing on several large and high-dimensional real datasets obtained from UCI machine repository. The experiments show that the MLDOF improves accuracy of anomaly detection with respect to LDOF and reduces the false positive error.

Keywords : Anomaly, LDOF, Index-based, Nested-Loop.

Contents

List of Figures

Chapter 1

Introduction

Outlier detection defined as finding outliers in data that do not show normal behavior.These data which do not conform are called as anomalies, outliers, exceptions. Anomaly and Outlier can be used interchangeably. Anomaly detection is important because it causes data translate to significant information in different variety of applications.

1.1 Outliers

An Outlier is an observation which is different from the others in sample. Outliers arise due to many reasons like malicious activity.Example credit card fraud etc. Outlier detection is different from noise removal though it is very much similar to it. A key difference between noise removal and outlier detection is data in noise removal that is

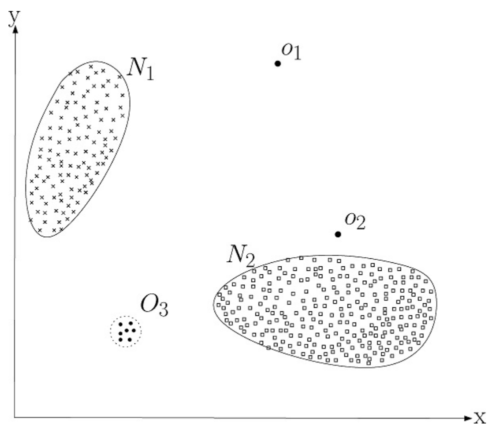

Figure 1.1: Outliers
N1 and N2 are normal data. o1 and o2 are two outliers and points in O3 is also considered as an outlier.

not of interest to analyst, but acts as a hindrance to data analysis whereas in outlier detection the data are interesting to the data analyst.

1.2 Causes of an Outliers

- Data quality poor / contaminated data
- Malfunctioning of equipment, manual error.
- Good but exceptional data

1.3 Outlier Detection Approach

It is used to define a area representing normal behavior and declare any area of observation in the data that does not belong to normal region as an outlier.

Some factors make this approach very challenging:

- Defining a normal region that encompasses every possible normal behavior is very difficult. The boundary between normal and anomalous behavior is often not precise. Anomalous observation that lies close to the boundary can be normal and vice versa.
- Due to malicious actions often anomalous observations appear normal, thereby making it difficult to define the normal behavior.
- The exact notion of an anomaly is different for different application domains. So appllication of a technique developed in one domain to a different domain cannot be straightforward.
- Major issue of models used by anomaly detection techniques is availability of labeled data for validating or training.
- The noise data tends to be similar to the actual anomalies and hence it is very difficult to distinguish and remove.

It is not easy to solve the problem of anomaly detection due to these challenges. In fact, most of the existing outlier detection techniques solve a specific formulation of the problem.

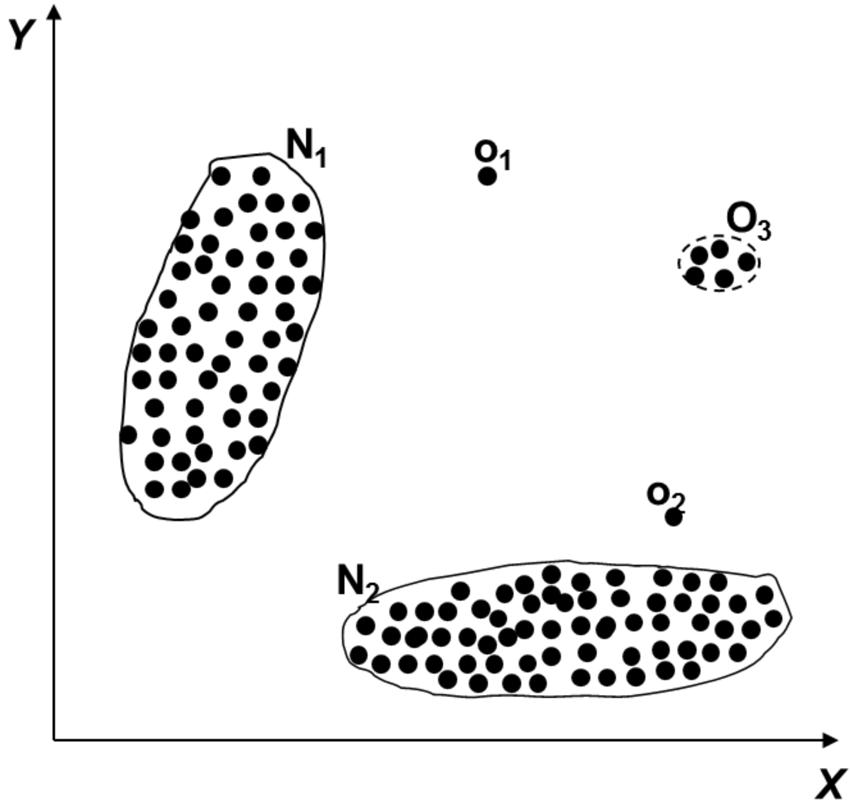

FIGURE 1.2: Example of anomalies in 2-D dataset

1.4 Outliers Types

The three categories of Anomalies classification are [1]:

1. Point Anomalies
   The instance is called outlier if it anomalous with respect to the rest of the data. It is simplest type of anomaly and is the focus of majority of research on anomaly detection.

2. Contextual Anomalies
   If a data instance is anomalous in a specific context but not otherwise. Induced by the structure in the dataset, as a part of the problem formulation the notion of a context has to be specified. Each data objects is defined using the following two sets of attributes:

   - Contextual attributes : These attributes are used to determine the context (or neighborhood) for that object. For object, in spatial datasets like determining the location, the latitude and longitude are the contextual attributes.

For a time-series data, the contextual attribute is the time which can determine the position of an instance on the entire sequence.

- Behavioral attributes : The behavioral attributes define the non-contextual characteristics of an instance. Considering a spatial dataset describing the average rainfall of the entire world. So at any location the amount of rainfall is the behavioral attribute.

3. Collective Anomalies: The collection of related data objects is anomalous with respect to the data is called collective anomalies.Individual data instances in a collective anomaly may not be outliers by themselves, but their occurrence together as a collection is anomalous.

1.5 Data Labels

The labels associated with data instances denote whether that instance is normal data or anomalous. Based on the extent to which the labels are available, the operation of anomaly detection can be done in one of the following three modes:

1.5.1 Supervised Anomaly Detection

Techniques trained in supervised mode assume that the availability of labeled instances of a training dataset for normal as well as anomaly classes. It's approach is to build a predictive model for normal against outlier classes. Any unseen data object is compared against the predictive model to determine which class it belongs to. The two major issues arise in supervised anomaly detection are discussed. Firstly , the anomalous instances are fewer compared to the normal instances in the training data. Due to imbalanced class distributions arise issues which have been addressed in the data mining and machine learning literature. Secondly, it is usually challenging to obtain accurate and representative labels for the anomaly class . There have been a number of techniques proposed that inject artificial anomalies into a normal dataset to obtain a labeled training dataset. Other than these two issues, the supervised anomaly detection problem is similar to building predictive models.

1.5.2 Semi-supervised Anomaly Detection

In this Techniques, it assume that the training data has labeled objects only for the normal class that operate in a semi-supervised mode. As they do not require labels for the anomaly class as they are more applicable than supervised techniques . For example, in spacecraft fault detection, it is not easy to model to signify an accident thats an anomaly scenario. The typical approach used in such techniques is to build a model for the class corresponding to its normal behavior and use this model to identify anomalies in the test data. A limited set of anomaly detection techniques exists that assumes availability of only the anomaly instances for training. As it is difficult to obtain a training dataset that covers every possible anomalous behavior that can occur in the data so such techniques are not commonly used.

1.5.3 Unsupervised Anomaly Detection

It do not require training data in those techniques that operate in unsupervised mode, and so most widely applicable. Assumption made that normal objects are far more frequent than outliers in test data in this category of techniques. Thus techniques suffer from high false positive rate if the assumption is not true. By using a sample of the unlabeled dataset as training data many semi-supervised techniques can be madde to operate in an unsupervised mode. IT assumes the test data contain very few anomalies and model learned during training is robust to few anomalies.

1.6 Types of Anomaly Detection Techniques

1. Statistical-based Detection
   (a) Distribution-based
   (b) Depth-based
2. Deviation-based Method
3. Sequential exception.
4. Distance-based Detection
   (a) Index-based
   (b) Nested-loop
   (c) Local-outliers
5. Density-based Detection
6. Clustering-based Detection

1.7 Output of Anomaly Detection

Anomaly detection techniques outputs are one of the following two types:

1. Scores : An anomaly score is assigned to each object in the test data according to their degree to which that object is considered an anomaly. The output of these techniques is a ranked list of anomalies. An analyst may choose to either analyze the top few anomalies or use a cutoff threshold to select the anomalies.
2. Labels : Each test object is assigned a label(normal or anomalous) using techniques in this category.

1.8 Challenges

- How many outliers are there in the data?
- Method is unsupervised Validation can be quite challenging (just like for clustering)
- Finding needle in a haystack.

1.9 Assumptions

Number of *normal* observation is much more than the number of *abnormal* observations(anomalies) in data.

1.10 Applications

1. Fraud detection
   Credit card fraud detection : Amount spent in a transaction which is very high compared to the normal transaction for that person,it will be a point anomaly.
2. Intrusion detection
   Network intrusion detection : Detection of anomalous activity (break-ins, penetrations, and other forms of computer abuse) in a computer related area.To gain unauthorized access for information form the network, hackers launch ciber attacks. It uses semi supervised and unsupervised anomaly detection techniques.

3. Fault detection

   Fault detection in Mechanical Units : Monitor the performance of industrial components such as oil flow in pipelines, turbines , motors and detect defects that might occur due to wear and tear or other unforeseen circumstances.

Chapter 2

Distance-based Outliers Methods

- Data is represented as a vector of features.
- It rely on the computation of distance value based on metric distance function.
- Distance-based outliers are rightly defined for k-dimensional datasets for any value of k[2].
- Distance-based: An object O in a dataset T is a DB(p,D) outlier if at least fraction p of the objects in T are grater than eual to distance D from O.
- The major approaches are
  - Nearest Neighbors based
  - Density based

2.1 Indexed-based Method

By analyzing the multidimensional indexing schemes [3] we found that, for variants of R-trees [4], k-d trees[5], andX-trees [6], range search lower bound complexity is (N11/k). By increasing the value of k, range search quickly reduces to O(N), giving at best a constant time improvement reflecting sequential search. So, the procedure for finding all DB(p, D) outliers has a worst case complexity of O(k N2). Compared to the depth-based approaches, which have a lower bound complexity of (N k 2), DB outliers scale much better with dimensionality. The framework of DBoutliers is applicable and computationally feasible for datasets that have many attributes (i.e., k 5). This is a significant improvement on the current state-of-the-art,where existing methods can only realistically deal with two attributes [7, 8].The above analysis only considers search time. When itcomes to using an index-based algorithm, most often for the kinds of data-mining applications under consideration, it is a very strong assumption that the right indexexists. In other words, a huge hidden cost to an index-based algorithm is the effort to build the index in the first place. As will be shown in Sect. 6, the index build-ing cost alone, even without counting the search cost,almost always renders the index-based algorithms non-competitive.

Steps

1. Index is used to search for neighbors of each object O within radius D around that object.
2. Once K (K = N(1-p)) neighbors of object O are found, O is not an outlier.
3. O(K*n2) is worst case computation complexity, K is the dimensionality and n is the no. of objects in the dataset[9].

Pros : Scale carefully with K
Cons : Index construction process may take much time.

2.2 Nested-loop Method

The nested loop calculate distances of instances to all other objects to find the instances k nearest neighbors. Nested loop has a complexity of O(kN2), (k is the no. of dimensions and N is the no. of data objects), and the no. of passes over dataset is linear to N. Due to quadratic complexity,in respect to no. of objects, making it deficient for mining large databases. Calculation of distances between objects is the major cost. Though Nested loop is a good choice for high dimensionality datasets, the large number of calculations makes it unsuitable[9].

Steps:

1. Divides the buffer space into two halves (first and second arrays)
2. Break data into blocks and then feed two blocks into the arrays.
3. Directly computes the distance between each pair of objects, inside the array or between arrays
4. Decide the outlier.
5. Same computational complexity as the index-based algorithm[10].
6. Pros:
   Avoid index structure construction
   Try to minimize the I/Os

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 outlier factor ,it 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 [11].

We use a test case in which there are 3 clusters C_1,C_2 and C_3 and four outlying points O_1,O_2,O_3,O_4 are present. 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. We solve this problem by following method .

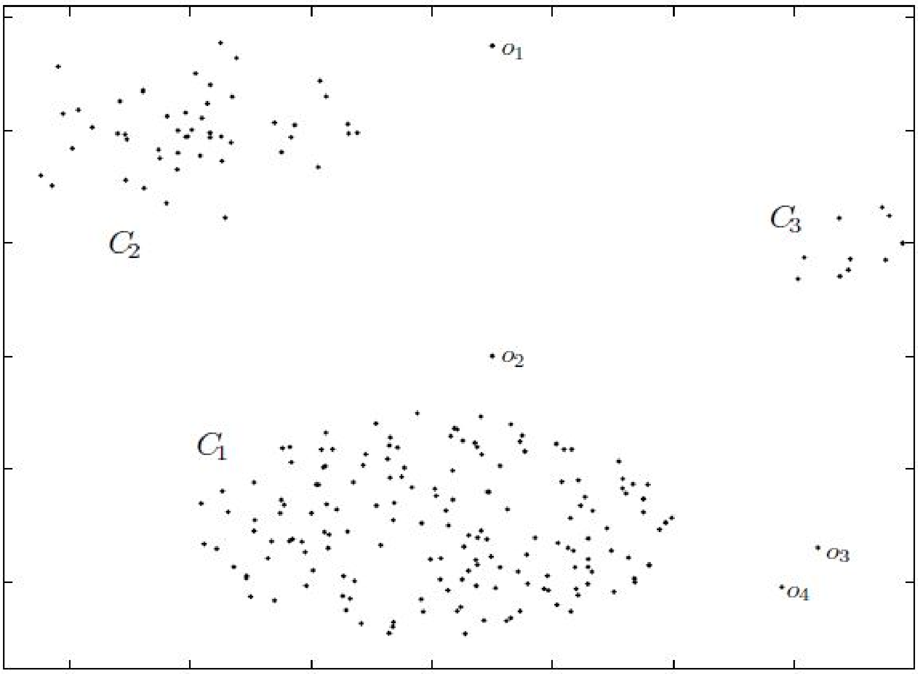

FIGURE 2.1: Test case to demonstrate effectiveness of LDOF

2.3.1 Formal Definitions

- **k^{th} inner 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 $\acute{o}$. The k-nearest neighbors distance of object x_p is defined as:
  $\bar{d}_{xp} = \frac{1}{k} \sum_{x_i \epsilon N_p} dist(x_i, x_p)$

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

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

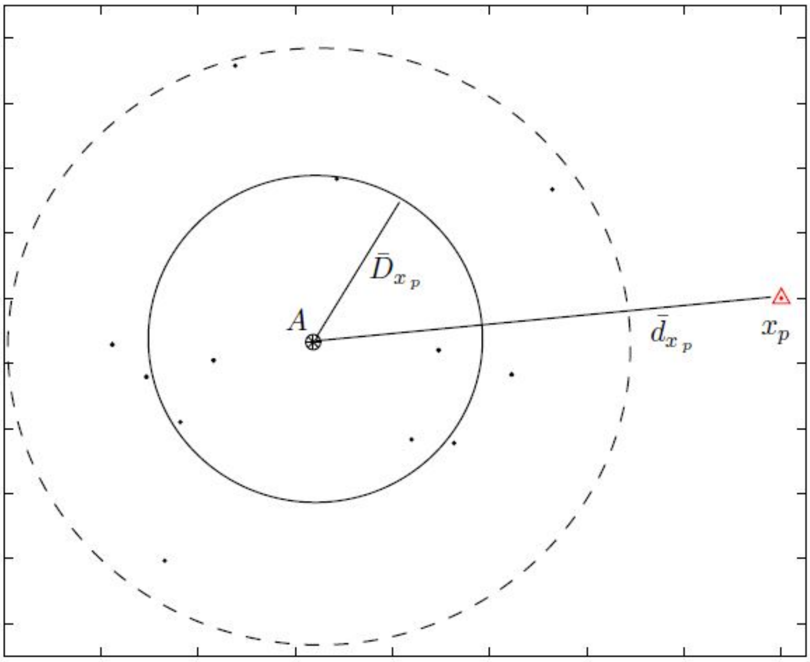

FIGURE 2.2: Showcasing $d_{x_p} and D_{x_p}$

Here $\bar{d}_{xp}$ mainly denotes the distance between the specified point and all other points in the k neighborhood set of the point and $\bar{D}_{xp}$ represents the distance between the points in the k neighborhood set[12]. Another way to minimize calculation is to find the median of the k neighborhood points and name is $\acute{x}$ and then calculate the 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 [11]:
Input : A given dataset D, natural numbers n and k.
Steps :

1. For each object o in D, retrieve o's k-nearest neighbors

2. Calculate LDOF for each object o. The objects with $LDOF < LDOF_{lb}$ are directly discarded

3. According to their LDOF values sort the object

Output : Highest LDOF values of first n objects.

Chapter 3

Motivation and Objective

3.1 Motivation

1. Solving the False Positive error of LDOF.

2. Improving the overall accuracy of outlier detection in numerical datasets

3.2 Objective

To improve the efficiency of existing LDOF algorithm by reducing false detection.

Chapter 4

Modified Local Distance-based Outlier Factor (MLDOF)

The disadvantage of Local Distance-based Outlier Factor is false detection i.e. false Positive error is high. We have introduced a new outlier detection method Modified Local Distance-Based Outlier Factor (MLDOF) to overcome and to improve the accuracy of the existing algorithm.

4.1 Features

1. It uses concept of KNN distance and KNN inner distance of a object.
2. It finds the Nearest neighbour of the points and decides it as outlier or not.
3. It reduces the false detection of normal data as outlier.

4.2 Algorithm

1. For each object p in D, retrieve p's k-nearest neighbours;
2. LDOF calulated for each object p. the objects having LDOF ¡ LDOFlb are discarded;
3. Objects sorted according to their LDOF values;
4. Identify indexes of point that can be a outlier.
5. Check for Nearest Neighbor for points that LDOF identified.

6. If M+1, neighbors are found in the D- neighborhood,
   then the search stops,
   and O declared as non-outlier.
   else
   O declared as outlier.

4.3 Explanation

In this algorithm, it takes each object form the dataset and computes k nearest neighbor individually. Nearest neighbor is calcluated using some underlying function. Then finds kNN distance with respect to the objects neighbors and also finds kNN inner distance of objects respectively. After that it computes the LDOF for each objects.If LDOF of an object is less than lower bound then it is discarded.After discarding then sort the objects according to their LDOF values. Then it identifies the indexes of point that can be a outlier. Now check for the nearest neighbors of that points that are identified by LDOF. If the nearest neighbor for points is greater than M where M=D(1-p)and p in minimum fraction of objects in dataset that must be outside the D-neighborhood of an outlier,then the points regarded as non-outlier. By doing that it discards some of the points identified by LDOF and that reduces ths false detection of normal data as an outliers.

Chapter 5

Experiment Details

In this algorithms,mainly uses k-nearest neighbor to find out the object is outlier or not.

5.1 Iris Data Set

Data Set Characteristics	Multivariate	**Area**	Life
Number of Instances	150	**Attribute Characteristics**	Real
Number of Attributes	4	**Associated Tasks**	Classification
Date Donated	1988-07-01	**Number of Web Hits**	447188

Attribute Information :

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

Source :

- Creator: R.A. Fisher
- Donor: Michael Marshall (MARSHALL
- Date: July, 1988

Information :

- This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda Hart, for example.) The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
- Predicted attribute: class of iris plant.
- This is an exceedingly simple domain.
- This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick@espeedaz.net) the 35th sample should be: 4.9, 3.1, 1.5, 0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9, 3.6, 1.4, 0.1,"Iris-setosa" where the errors are in the second and third features.

5.2 Haberman's Survival Data Set

Data Set Characteristics	Multivariate	**Area**	Life
Number of Instances	306	**Attribute Characteristics**	Integer
Number of Attributes	3	**Associated Tasks**	Classification
Date Donated	1999-03-04	**Number of Web Hits**	47250

Attribute Information :

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
   1 = the patient survived 5 years or longer
   2 = the patient died within 5 year

Source :

- Donor: Tjen-Sien Lim (limt '@' stat.wisc.edu)

Information :

Dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

5.3 Seeds Data Set

Data Set Characteristics	Multivariate	**Area**	Life
Number of Instances	210	**Attribute Characteristics**	Integer
Number of Attributes	7	**Associated Tasks**	Classification, Clustering
Date Donated	2012-09-29	**Number of Web Hits**	22190

Attribute Information :

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

1. Different varieties of wheat:
   (a) Kama
   (b) Rosa
   (c) Canadian
2. Area A
3. Perimeter P
4. Length of kernel
5. Width of kernel
6. Asymmetry coefficient
7. Compactness C = $4*pi*A/P^2$

Source :

- Magorzata Charytanowicz, Jerzy Niewczas Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, Konstantynw 1 H, PL 20-708 Lublin, Poland. E-mail: mchmat,jniewczas@kul.lublin.pl
- Piotr Kulczycki, Piotr A. Kowalski, Szymon Lukasik, Slawomir Zak Department of Automatic Control and Information Technology, Cracow University of Technology, Warszawska 24, PL 31-155 Cracow, Poland and Systems Research Institute, Polish Academy of Sciences, Newelska 6, PL 01-447 Warsaw, Poland.E-mail: kulczycki,pakowal,slukasik,slzak@ibspan.waw.pl

Information :

- The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment.
- High quality visualization of the internal kernel structure was detected using a soft X-ray technique.
- It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology.
- The images were recorded on 13x18 cm X-ray KODAK plates.
- Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.
- The dataset can be used for the tasks of classification and cluster analysis.

5.4 Breast Cancer Wisconsin (Original) Data Set

Data Set Characteristics	Multivariate	**Area**	Life
Number of Instances	699	**Attribute Characteristics**	Integer
Number of Attributes	10	**Associated Tasks**	Classification, Clustering
Date Donated	1992-07-15	**Number of Web Hits**	90885

Attribute Information :

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10

11. Class:
    2 for benign
    4 for malignant)

Source :

- Creator: Dr. WIlliam H. Wolberg (physician),University of Wisconsin Hospitals Madison, Wisconsin, USA
- Donor: Olvi Mangasarian (mangasarian '@' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)

Information :

- Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself :
  1. Group 1: 367 instances (January 1989)
  2. Group 2: 70 instances (October 1989)
  3. Group 3: 31 instances (February 1990)
  4. Group 4: 17 instances (April 1990)
  5. Group 5: 48 instances (August 1990)
  6. Group 6: 49 instances (Updated January 1991)
  7. Group 7: 31 instances (June 1991)
  8. Group 8: 86 instances (November 1991)
  9. Total: 699 points (as of the donated datbase on 15 July 1992)
- Note that the results summarized above in Past Usage refer to a dataset of size 369, while Group 1 has only 367 instances. This is because it originally contained 369 instances; 2 were removed.

Chapter 6

Results and Discussions

We implemented both LDOF and MLDOF outlier detection methods on the datasets over a large range of k.

In a classification task, the precision for a class is no. of true positives divided by the total no. of elements labeled as belonging to the positive class. A perfect precision score of 1 means that every result obtained by a search was relevant.
Example: A precision value of 1 for a class X means that every item labeled as belonging to class X does indeed belong to class X.
Here precision is calculated to evaluate the performance of each algorithm.

Precision= N_D / N_R

where,
N_D= Number of detected real outliers in the dataset
N_R=Number of real outliers in the dataset.

6.1 Iris Dataset

We have taken total 60 instances in which 50 instances are normal and 10 instances are outliers. Knowingly 10 outliers are taken to test the algorithms are working fine or not. We ran both algorithms, for k=12; LDOF showed 13 instances as outliers where 3 instances are normal data which falsely detected as outliers and other 10 instances are real outliers but in MLDOF showed 10 instances which are real outliers. We draw confusion matrix for both the algorithms in figure.6.1 and figure.6.2. In order to compute performance of the algorithm we draw the precision vs neighborhood size k graph, it is shown in figure 6.3.

1. Input 50 instances of a class 10 instances of other class
2. Output:The following indexes as outliers [51 52 53 54 55 56 57 58 59 60]

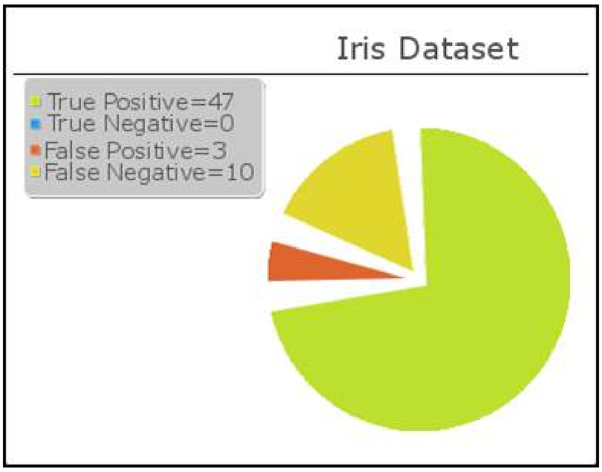

Figure 6.1: Output of LDOF in Iris dataset

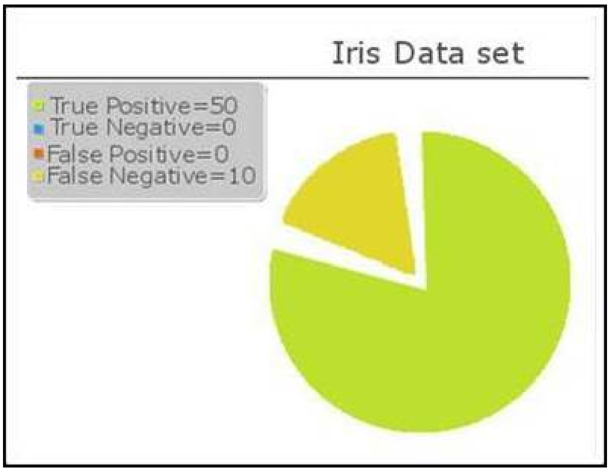

Figure 6.2: Output of MLDOF in Iris dataset

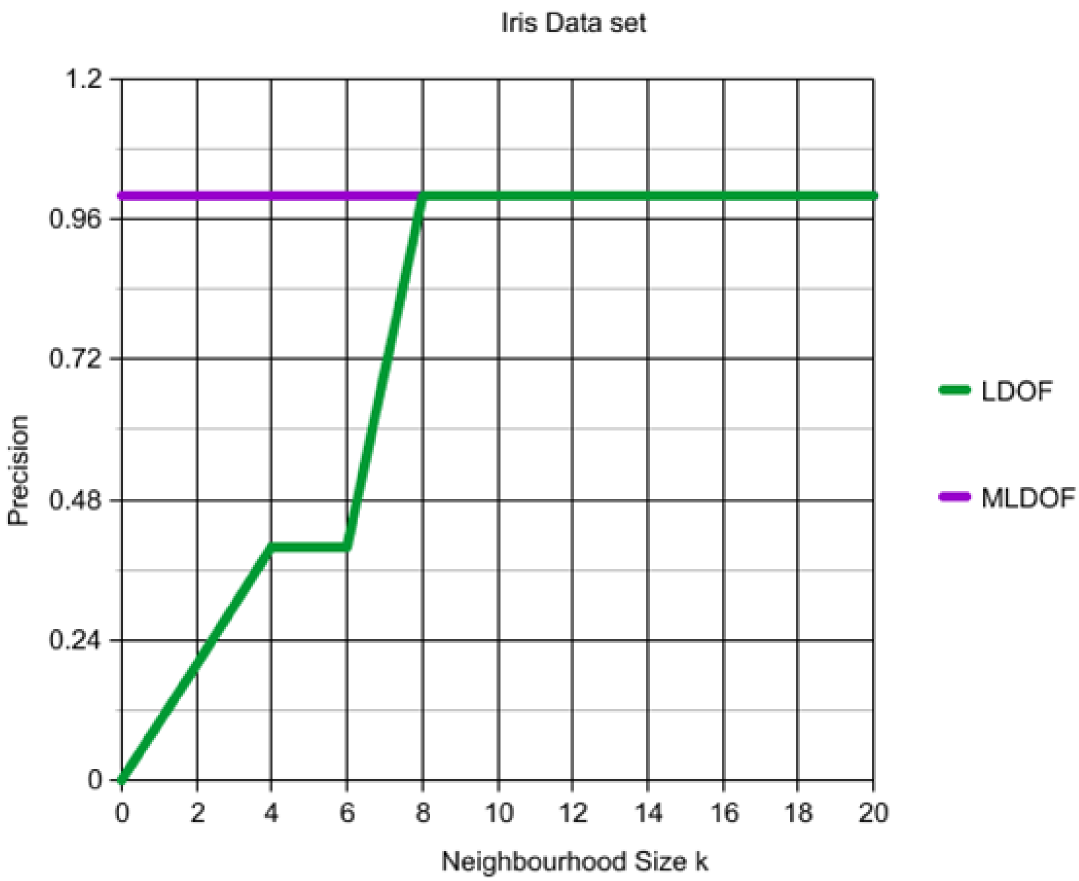

Figure 6.3: Precision in Iris dataset

6.2 Haberman's Survival Dataset

We have taken total 229 instances in which 225 instances are normal and 4 instances are outliers.Knowingly 4 outliers are taken to test the algorithms are working fine or not. We ran both algorithms, for k=6; LDOF showed 11 instances as outliers where 7 instances are normal data which falsely detected as outliers and other 10 instances are real outliers but in MLDOF showed 5 instances in which 4 are real outliers and 1 normal data . We draw confusion matrix for both the algorithms in figure.6.4 and figure.6.5. In order to compute performance of the algorithm we draw the precision vs neighborhood size k graph, it is shown in figure 6.7. The 3-D projection of the dataset is drawn showing normal data and outliers.

1. Input: 225 instances of a class 4 instances of other class
2. The following indexes as outliers [226 227 228 229]

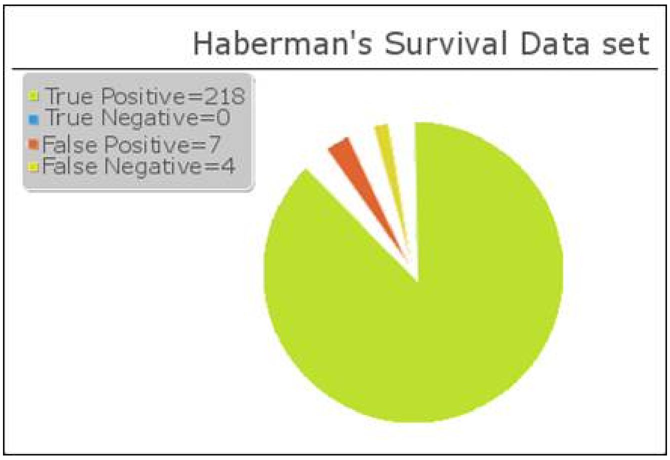

FIGURE 6.4: Output of LDOF in Haberman's Survival dataset

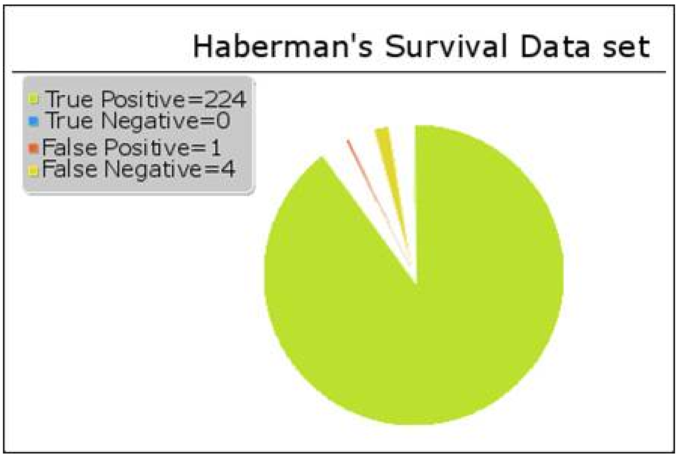

FIGURE 6.5: Output of MLDOF in Haberman's Survival dataset

6.3 Seed Dataset

We have taken total 150 instances in which 140 instances are normal and 10 instances are outliers. Knowingly 10 outliers are taken to test the algorithms are working fine or not. We ran both algorithms, for k=12; LDOF showed 14 instances as outliers where 4 instances are normal data which falsely detected as outliers and other 10 instances are real outliers but in MLDOF showed 10 instances which are real outliers. We draw confusion matrix for both the algorithms in figure.6.8 and figure.6.9. In order to compute performance of the algorithm we draw the precision vs neighborhood size k graph, it is shown in figure 6.10.

1. Input: 70 instances of 2 classes each 10 instances of other class
2. The following indexes as outliers [141 142 143 144 145 146 147 148 149 150]

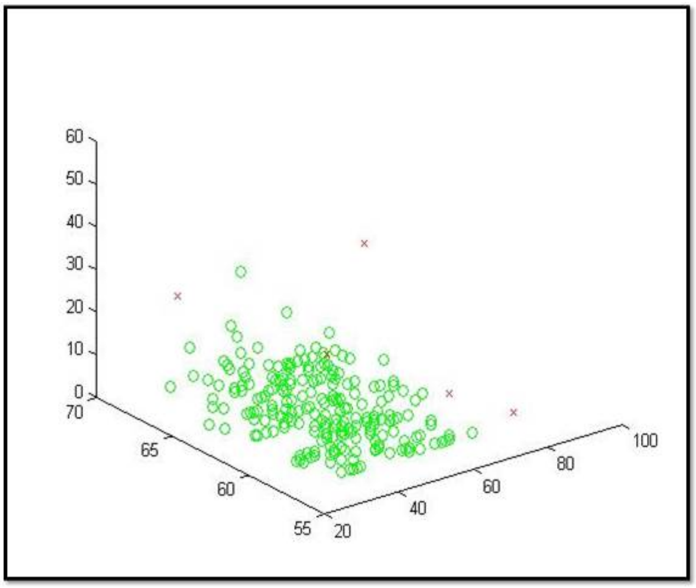

Figure 6.6: The 3-D projection of Haberman's Survival Dataset

6.4 Breast Cancer Dataset

We have taken total 449 instances in which 444 instances are normal and 5 instances are outliers. Knowingly 5 outliers are taken to test the algorithms are working fine or not. We ran both algorithms, for k=10; LDOF showed 11 instances as outliers where 6 instances are normal data which falsely detected as outliers and other 5 instances are real outliers but in MLDOF showed 7 instances in which 5 are real outliersand 2 normal data. We draw confusion matrix for both the algorithms in figure.6.11 and figure.6.12.In order to compute performance of the algorithm we drawn the precision vs neighborhood size k graph, it is shown in figure 6.13.

1. Input: 444 instances of a class 5 instances of other class
2. Output: The following indexes as outliers [445 446 447 448 449]

FIGURE 6.7: Precision in Haberman's Survival dataset

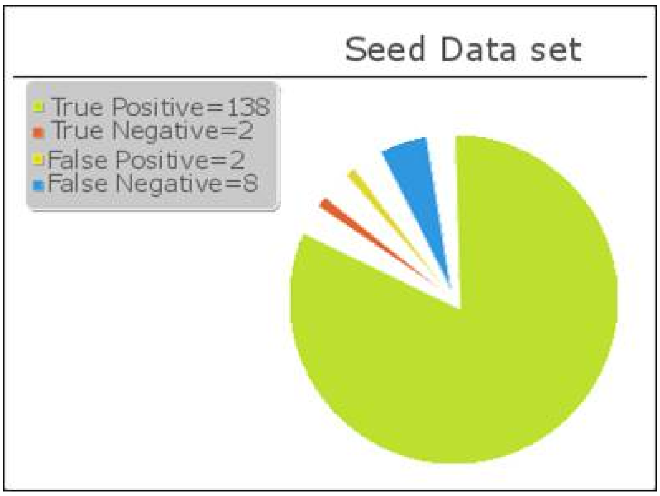

FIGURE 6.8: Output of LDOF in Seed dataset

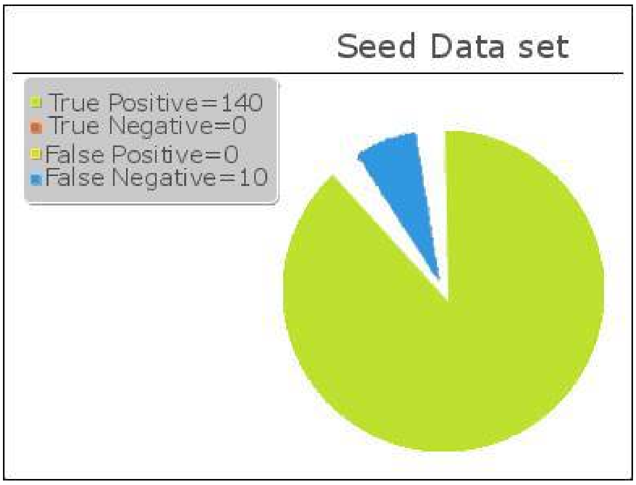

FIGURE 6.9: Output of MLDOF in Seed dataset

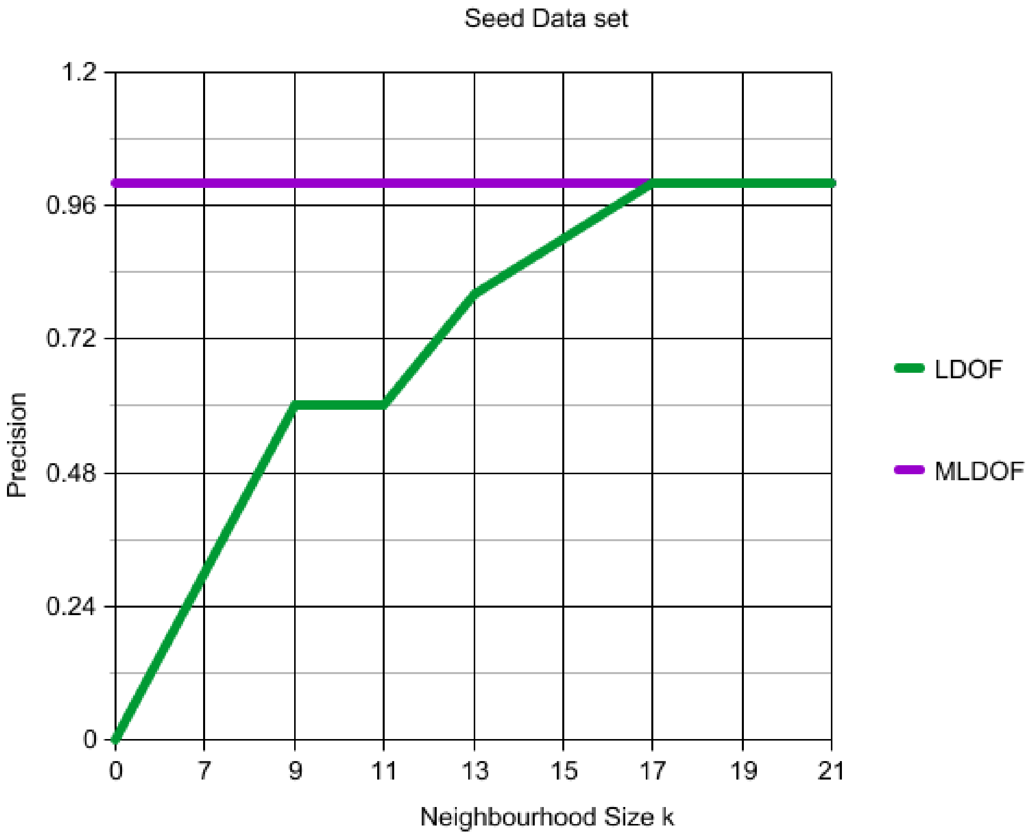

FIGURE 6.10: Precision in Seed dataset

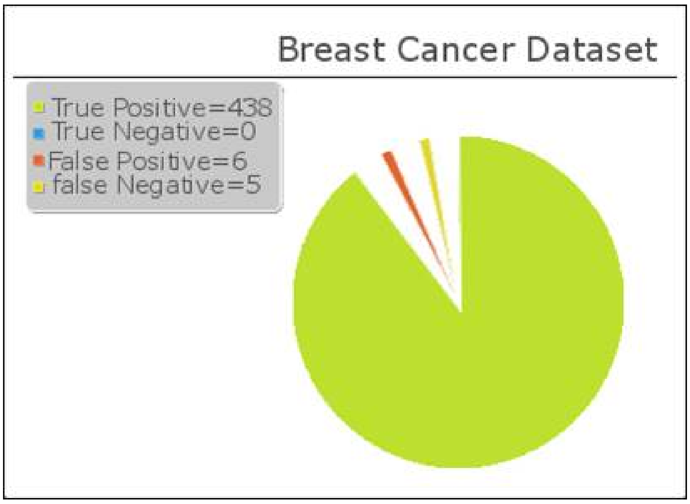

Figure 6.11: Output of LDOF in Breast Cancer dataset

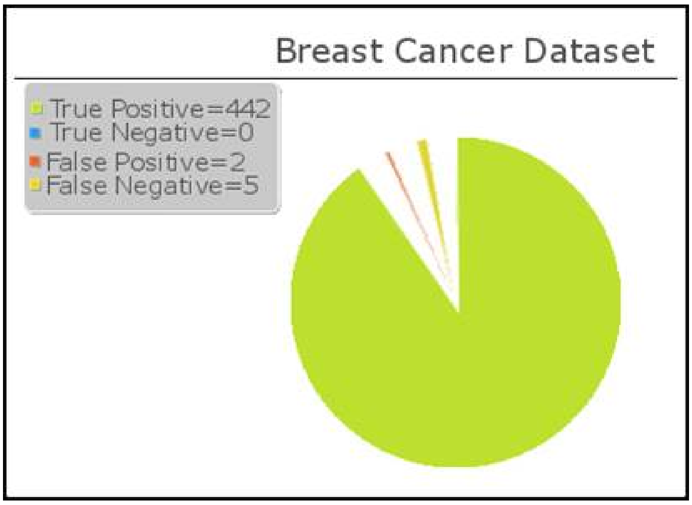

Figure 6.12: Output of MLDOF in Breast Cancer dataset

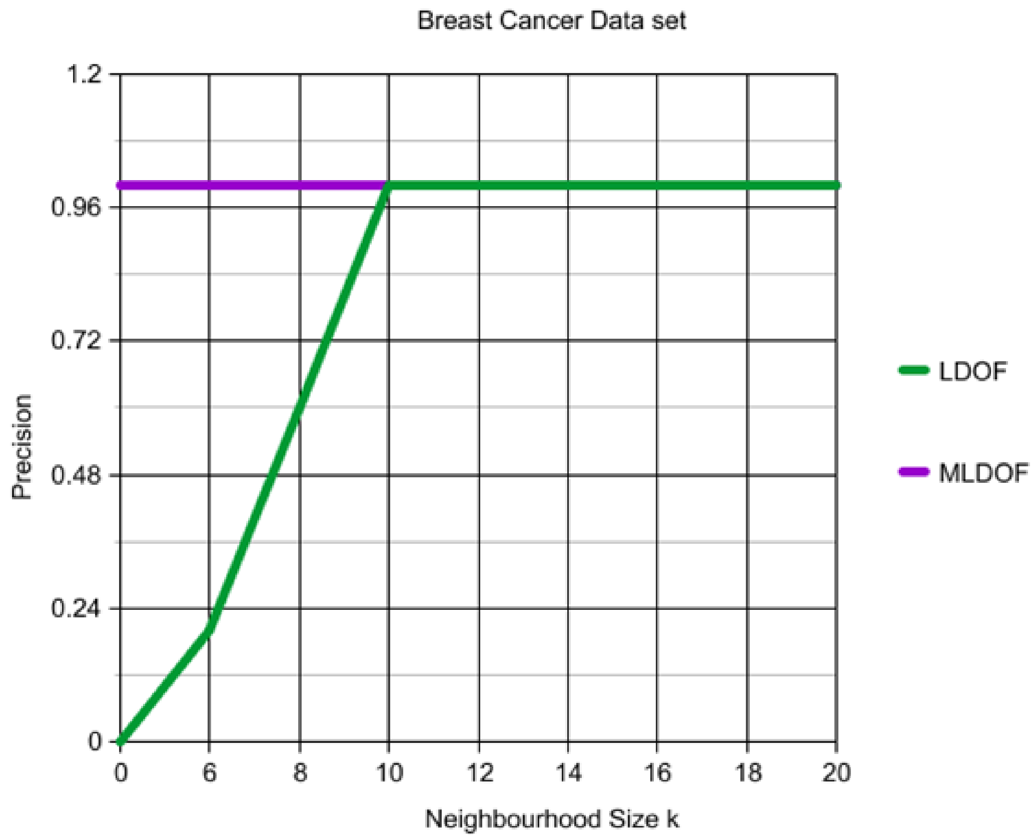

Figure 6.13: Precision in Breast Cancer dataset

Chapter 7

Conclusion

In this thesis we have introduced a new outlier detection method Modified Local Distance-Based Outlier Factor (MLDOF).Though K-nearest neighbor algorithm is easy to implement but to determine the value of K at prior is a diffcult task.Many number of iterations were performed to determine the optimal value of K which was a very time consuming process. We have shown that the Modified Local Distance Outlier Factor (MLDOF) improves the accuracy of outlier detection with respect to LDOF and reduces the False Positive Error.

Bibliography

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[2] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, volume VLDB '98, pages 392–403. Citeseer, 1998.

[3] Joseph M Hellerstein, Elias Koutsoupias, and Christos H Papadimitriou. On the analysis of indexing schemes. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 249–256. ACM, 1997.

[4] Antonin Guttman. *R-trees: a dynamic index structure for spatial searching*, volume 14. ACM, 1984.

[5] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[6] Stefan Berchtold, Daniel A Keim, and Hans-Peter Kriegel. The x-tree: An index structure for high-dimensional data. *Readings in multimedia computing and networking*, page 451, 2001.

[7] Theodore Johnson, Ivy Kwok, and Raymond Ng. Fast computation of 2-dimensional depth contours. In *Proc. KDD*, volume 1998, pages 224–228. Citeseer, 1998.

[8] Ida Ruts and Peter J Rousseeuw. Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, 23(1):153–168, 1996.

[9] Edward Hung and David W Cheung. Parallel mining of outliers in large database. *Distributed and Parallel Databases*, 12(1):5–26, 2002.

[10] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, volume 29, pages 427–438. ACM, 2000.

[11] Ke Zhang, Marcus Hutter, and Huidong Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Advances in Knowledge Discovery and Data Mining*, pages 813–822. Springer, 2009.

[12] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pages 93–104. ACM, 2000.