DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

UNSUPERVISED ANOMALY DETECTION ALGORITHMS

by
Beyza KIZILKAYA

January, 2019
İZMİR

UNSUPERVISED ANOMALY DETECTION ALGORITHMS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In partial Fulfilment of the Requirements for the Master of Science in
Statistics, Statistic Program

by

Beyza KIZILKAYA

January, 2019
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"UNSUPERVISED ANOMALY DETECTION ALGORITHMS"** completed by **BEYZA KIZILKAYA** under supervision of **ASST. PROF. DR. ENGİN YILDIZTEPE** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Engin YILDIZTEPE

Supervisor

Assoc. Prof. Dr. Hatice Savaş SAZAK

(Jury Member)

Assoc. Prof. Dr. A. Fırat Özdemir

(Jury Member)

Prof.Dr. Kadriye ERTEKİN
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude and special appreciation to my advisor, Asst. Prof. Dr. Engin YILDIZTEPE, for his excellent and continuous support, patience and encouragement throughout my thesis. His guidance and experiences played very important role in my all research.

I would like to acknowledge my dissertation committee members, Assoc. Prof. Dr. Abdullah Fırat ÖZDEMİR and Assoc. Prof. Dr. Hakan Savaş SAZAK, for their constructive comments and suggestions.

I would like to thank to my family for their excellent support. I would like to express my special thanks to my dear, lovely husband Alp Cem KIZILKAYA, for his limitless patience and support during this thesis.

Beyza KIZILKAYA

UNSUPERVISED ANOMALY DETECTION ALGORITHMS

ABSTRACT

Detection of outliers or anomalies in the data is of great importance in data analysis. Different approaches can be used in anomaly detection according to type of the problem. Unsupervised anomaly detection (UAD) approach is the most challengeable part of these approaches. UAD methods aim to detect anomalies without using a labelled training dataset. UAD algorithms can be considered in three main groups: nearest neighbour based, clustering based and statistical based. In this thesis, UAD approaches is examined and an adjustment, that depends on sample size, is proposed for statistical based algorithm, HBOS. In the first part of the application, performance of the most widely used UAD algorithms, that are k-nearest neighbour (k-NN), local outlier factor (LOF), local density cluster-based outlier factor (LDCOF) and histogram-based outlier score (HBOS), are compared. According to the results, HBOS algorithm is found more successful in terms of accuracy rate and runtime. In the second part of the application, effect of the bin-width determination techniques on to the performance of HBOS algorithm is examined. According to the results of the comparison with multivariate data of different characteristics, there is no superiority between the bin-width determination techniques in terms of accuracy.

Keywords: Anomaly, unsupervised anomaly detection, local outlier factor, local density cluster based outlier factor, histogram based outlier score

DENETİMSİZ ANOMALİ TESPİT ALGORİTMALARI

ÖZ

Verilerdeki aykırı değerlerin veya diğer bir deyişle anomalilerin tespit edilmesi veri analizinde büyük önem taşımaktadır. Anomali tespitinde probleme göre farklı yaklaşımlar kullanılabilir. Denetimsiz anomali tespiti (DAT), bu yaklaşımlar içerisinde en zorlu olanıdır. DAT yöntemleri etiketli bir eğitim kümesi kullanmadan anomalileri belirlemeyi hedefler. DAT algoritmaları, en yakın komşuluk tabanlı, kümeleme tabanlı ve istatistiksel tabanlı olmak üzere üç ana başlık altında incelenebilir. Bu tezde, DAT yaklaşımı incelenmiş ve istatistiksel tabanlı HBOS algoritması için örneklem genişliğine duyarlı bir düzeltme önerisinde bulunulmuştur. Uygulamanın ilk bölümünde yaygın olarak kullanılan, k en yakın komşu (k-NN), yerel aykırı değer (LOF), yerel yoğunluk kümeleme aykırı değer (LDCOF), ve histogram tabanlı aykırı değer skoru (HBOS) algoritmalarının performansları karşılaştırılmıştır. Elde edilen sonuçlara göre, HBOS algoritması doğruluk oranı ve çalışma süresi açısından daha başarılı bulunmuştur. Uygulamanın ikinci bölümünde farklı kutu genişliği belirleme tekniklerinin HBOS algoritmasının performansına etkileri araştırılmıştır. Farklı özelliklerde seçilen çok değişkenli veriler ile yapılan karşılaştırma sonuçlarına göre doğruluk oranı açısından kutu genişliği belirleme teknikleri arasında üstünlük saptanamamıştır.

Anahtar kelimeler: Anomali, denetimsiz anomali tespiti, yerel aykırı değer, histogram tabanlı aykırı değer skoru

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Anomaly detection is an important part of data analysis which provides correct and healthy evaluation of the data. It's the mostly about finding uncomfortable patterns in the rest of the data (Chandola, Banerjee, & Kumar, 2009). In the literature, there are many definitions for anomalies such as outliers, novelties, exceptions, surprises, discordant observations. These terms are similar, and they can be used interchangeably. Statistic community deals with the outliers since 19^{th} century. Outlier term is one of the most used terms in literature that is used instead of anomaly term and actually there is no difference between these two terms, because in fact, outliers are point anomalies. While point anomalies may be named as outliers, there are two different type of anomaly which are defined in the literature other than point anomaly. The first one, contextual anomaly is seen in time series data mostly. The latter, collective anomaly defines a group of observations that are anomaly together, but these observations are not anomaly individually.

Statistical learning is about understanding data and it has two main categories as supervised and unsupervised learning (James, Daniela, Hastie, & Tibshirani, 2015). Anomaly detection approaches come from the learning idea and there are two main categories of anomaly detection approaches which are supervised anomaly detection and unsupervised anomaly detection. These two concepts have a main difference in terms of applications. Supervised approaches need labelled data for both normal and anomaly classes, but unsupervised approaches don't need any labelled data, they only learn from data. Recently, a third approach, semi-supervised anomaly detection has been developed for anomaly detection approach which is a hybrid of supervised and unsupervised approaches. Semi-supervised methods need labelled data only for normal classes.

Unsupervised anomaly detection is about finding anomalous patterns that are accepted as anomaly with respect to the rest of the data (Guthrie, Guthrie, Allison, & Wilks, 2007). Unsupervised anomaly detection is considered to be the most difficult

approach to detect anomaly because no labelled training dataset is used. There are three main categories of unsupervised anomaly detection approach; nearest-neighbour based, cluster-based, and statistical-based. Many unsupervised algorithms have been developed under these three main unsupervised anomaly detection categories.

In this thesis, anomaly, anomaly types, and anomaly detection approaches are defined and four unsupervised anomaly detection techniques are investigated. In the application, performances of the four unsupervised anomaly detection algorithms are compared using 14 real and 2 artificial datasets from different application domains such as health-care systems, intrusion detection, and image processing. Both univariate and multivariate datasets are used in the applications. Also, a statistical based unsupervised anomaly detection algorithm Histogram-Based Outlier Score (HBOS) is investigated in terms of histogram bin-width determination techniques.

This thesis is organized in five chapters. Anomaly detection concept, anomaly types, anomaly detection approaches are described in chapter two, unsupervised anomaly detection methods are given in chapter three, the applications and their results are given in chapter four. Finally, concluding remarks and some suggestions about future studies are given in chapter five.

CHAPTER TWO
ANOMALY DETECTION

Many methods used to identify anomalies are basically similar but given different names. In this study, anomaly detection term has been chosen to call the task. Anomaly detection is one of the most important tasks in data analysis. Finding abnormal instances in data is used in the variety of applications such as fraud detection, health care, intrusion detection, and surveillance systems. Anomaly determination is also an active research area in different disciplines such as statistics, machine learning, and computer science. In this chapter, anomaly, anomaly types, and anomaly detection approaches are described.

2.1 Definition of Anomaly

Anomaly is an unwanted, surprise pattern or sometimes a data point in a dataset. This pattern or data point differs from rest of the data with the abnormal behaviour. Different terms can be used in literature for anomaly such as outlier, discordant observation, and exception (Chandola et al., 2009). These terms can be used interchangeably, but there may be little differences between them for application domains. Anomalies can arise because of a user's mistake in data entry or may be a real unwanted observation in a dataset. Sometimes anomalies can arise as a result of a malicious attack like network hacking. Malicious anomalies behave proper to dataset and it's hard to finding them in data stream.

2.2 Anomaly Types

There are key components in detection of anomaly process and one of these key components is the nature of data. Data row in a dataset is named as record, observation, instance, and object. Inputs are defined as a collection of a data instance and data instances are defined as attributes or sometimes as variables, features, dimensions. Attributes can be binary, categorical or continuous. Input type has effect on anomaly type and anomaly detection algorithm type. Another important key component of

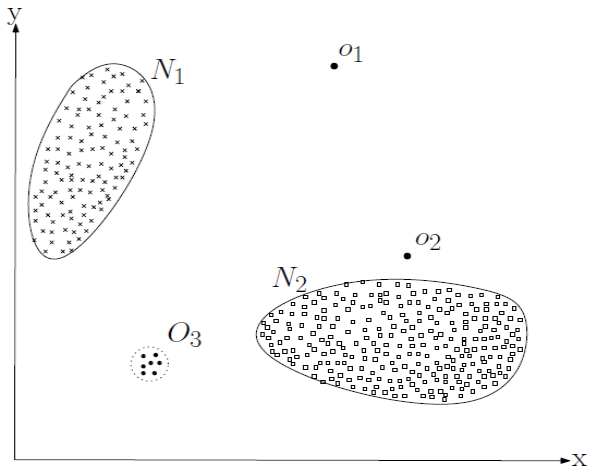
anomaly detection process is anomaly type. As input type of data instance, type of anomaly has effect on anomaly detection method that will be applied. Anomalies can be classified into three categories. The graphical examples of the anomaly types are given in Figure-1.

2.2.1 Point Anomalies

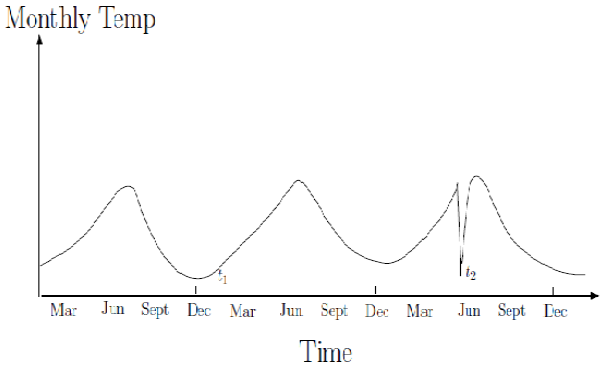
Point anomalies are the simplest type of anomalies. Point anomaly is a single data instance which differs from the remainder of the data. An unexpected observation may be evaluated as a point anomaly. For example, a sudden change in the usage of the credit card may be fraudulent usage such as stolen credit card. A transaction amount which is very higher than usual spends is a point anomaly.

2.2.2 Contextual Anomalies

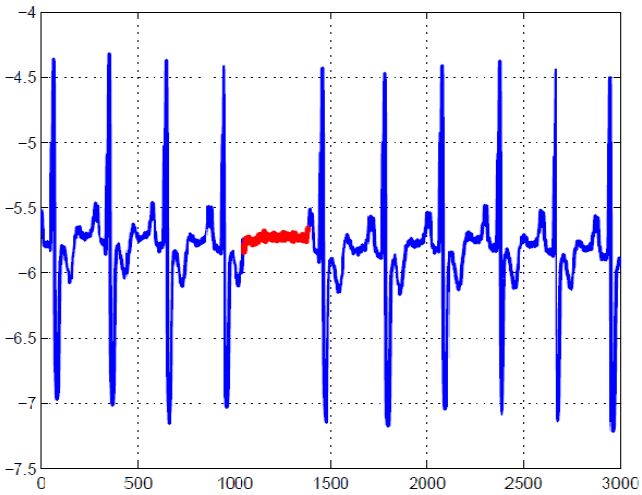
Contextual anomaly depends on given context. The same data instances could be considered anomalous in a specific context and normal in another. Anomalies in time series data is a good example for contextual anomalies. Because, this kind of anomalies are accepted as an anomaly if they are only in the specific time, otherwise, they may be accepted normal data instances. For that reason, contextual anomalies are also called “Conditional Anomalies”. Increasing on ice-cream sales in summer is accepted as a normal behaviour, but increasing on sales in winter isn't expected, and in this specific context, increasing on ice-cream sales in winter is accepted as a contextual anomaly.



Point Anomaly



Contextual Anomaly



Collective Anomaly

Figure 2.1 Anomaly types (Chandola et al., 2009)

2.2.3 Collective Anomalies

Collective anomalies are a group of observations and these points are not anomaly individually, they are accepted as an anomaly if and only if they occur together as a collection. Human electrocardiogram result or web attacks are good examples of this kind anomalies.

In addition, anomalies are evaluated according to two approaches that are “Global Anomalies” and “Local Anomalies”. Global anomalies are clearly different from the dense areas in the data. Local anomalies are considered anomalous compared to their neighbours. Local anomaly methods can detect anomalies which cannot be detected by global anomalies.

2.3 Anomaly Detection

Anomaly detection is about finding nonconforming patterns in a dataset. These patterns are seen very rare in a dataset and in most cases, it’s hard to distinct their behaviours from normal instances. Anomaly detection is used in many application domains such as intrusion detection, fraud detection, health-care systems, damage detection systems. Also, there is a wide research area of anomaly detection concept such as statistics, machine learning, data mining, information theory and so on. All these research areas deal with anomaly detection in different aspects; for example, a statistician may examine anomaly detection in terms of eliminating outliers.

As it’s defined before anomaly detection is about finding anomalous patterns in the data. But this is the simplest thinking way for anomaly detection. In real, anomaly detection has not a straightforward way. There are many challenges to distinguish anomalies from normal instances:

- Normal behaviour may be defined differently in different application domains; for example; a minor anomaly point may have no effect on the financial markets, but the same size anomaly point has major importance on a diagnose of an important disease.

- The defining frame of normal behaviour is hard in terms of many ways. Difference between normal and abnormal behaviour may not be exact every time. That's why the boundaries must be drawn very carefully to define the difference between normal and abnormal behaviour.
- If anomalies are results of a malicious attack, then anomaly points can adapt themselves to the normal instances.
- Definition of normal behaviour may change in time for application domains. That's why the definition of normal behaviour related application domains may be expired in later times.

Many challenge factors can be added to the above list for anomaly detection. Also, every research area and application domains have different challenge factors for anomaly detection. These challenges show that the solution of an anomaly detection problem is not easy. That's why research areas and application domains work in a collaboration about anomaly detection.

2.4 Related Studies

In this section, some studies that use or compare anomaly detection algorithms are included. The research by Eskin et al. (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002) is an example of using unsupervised anomaly detection in intrusion detection. In 2005, Leung and Leckie published a study about anomaly detection for network intrusion detection (Leung & Leckie, 2005). One of the most popular application domains recently is machine learning. Tsai et al. published a review study of anomaly detection for intrusion detection systems for machine learning in 2009 (Tsai, Hsu, Lin, & Lin, 2009).

Intrusion detection is not only application domain for anomaly detection studies. Chuah and Fu studied about medical sensors to detect anomalies in ECG observations in 2007 (Chuah & Fu, 2007). Oh and Shon studied damage diagnosis on health care systems with using support vector machine algorithm in 2009 (Oh & Shon, 2009). One of the interesting studies about anomaly detection was published in 2009. Schwabacher et al. applied unsupervised anomaly detection algorithms to data that were collected

from two rocket propulsion testbeds (Schwabacher, Oza, & Matthews, 2009). Bontemps et al. published a study in 2016 and proposed a method to detect collective anomalies for neural network long short-term memory (Bontemps, McDermott, & Le-Khac, 2016). One of the recently striking studies was about understanding the behaviour of bees (Gama, Arruda, Carvalho, Souza, & Pessin, 2017). They used the local outlier factor algorithm to evaluate models for the classification of anomalous events. A comparison study by Goldstein and Uchida give a comprehensive evaluation of unsupervised anomaly detection algorithms (Goldstein & Uchida, 2016).

A detailed review of outlier detection techniques was published by Austin and Hodge in 2004 (Hodge & Austin, 2004). The most recent review study was published by Chandola. (Chandola et al., 2009) in 2009. These two studies present the state of art of the anomaly detection concept.

2.5 Anomaly Detection Output

Anomaly detection algorithms can produce two possible outcomes. Scoring algorithms calculate an anomaly score for every data instance. This score can be considered as a degree of outlierness. The anomaly score can be converted to the anomaly flag by using a particular threshold value. The output of the labelling algorithms is a binary result. They can only assign a label to the data instance as normal or abnormal. Labelling algorithms are more general than the scoring. But scoring lets the researchers make comment relative anomalies using a threshold value.

2.6 Anomaly Detection Approaches

Choosing the correct anomaly detection method according to the application's purpose, data, and type of output is very critical for the process. In this step, firstly the anomaly determination approach should be chosen. The anomaly detection methods are categorized into three main approaches as shown in Figure–2. These approaches are described in the following sections.

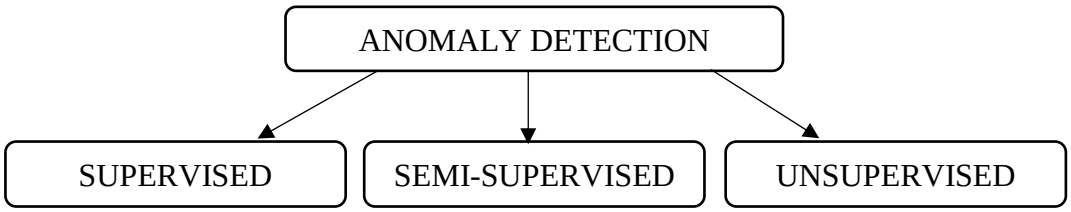


Figure 2.2 Anomaly detection approaches

2.6.1 Supervised Anomaly Detection

Supervised anomaly detection methods requires training dataset. The data instances in the training data set must have a label indicating whether they are anomalous. Supervised algorithms learn from the training dataset. Most of the classification based anomaly detection methods are in this category. Training data to be used in the classification algorithm should include a good distribution of both normal and abnormal samples to allow generalization by the classifier (Hodge & Austin, 2004).

The use of supervised anomaly detection methods has very important advantages and disadvantages. One of the most important advantages of these methods is that they provide reliable results when using the appropriate training data set. Another important advantage is the testing process. With the test data, the success of the model can be predicted. The major disadvantage of supervised methods is the high cost (Hodge & Austin, 2004). Because it is time consuming and difficult to obtain whether each data instance in the training data is anomaly. In some applications this may not be possible. Another disadvantage is that labelling is sometimes insufficient to detect anomalies. For example, local anomalies may not be detected only by labelling, they also need to be scored for detection (Goldstein & Uchida, 2016)

2.6.2 Semi-Supervised Anomaly Detection

Semi-supervised anomaly detection techniques are hybrid of supervised and unsupervised techniques. Semi-supervised methods use labelled data for only normal classes, there is no need to anomaly classes. Because of this specialty they are more applicable then the supervised anomaly detection techniques. Semi-supervised

techniques use unsupervised algorithms mostly to detect anomalies in a dataset. Training set is labelled for only normal data instances, and a model is built to predict anomaly instances from training dataset. Semi-supervised methods are proper to evolve to unsupervised methods as of structure. But these type methods are not commonly preferred to use for anomaly detection tasks (Chandola et al., 2009).

2.6.3 Unsupervised Anomaly Detection

Unsupervised anomaly detection is the hardest part of anomaly detection concept. Contrary to supervised or semi-supervised methods datasets are not labelled as normal and anomaly classes in unsupervised anomaly detection methods, that's why, performance of algorithm to detect anomalies have more important than the other approaches. Unlabelling process puts unsupervised anomaly detection algorithms to the most striking category of anomaly detection concept, with this feature of unsupervised anomaly detection algorithms, researchers have developed many algorithms at this area. These algorithms are in three main categories: Nearest-neighbour based algorithms, cluster-based algorithms and statistical-based algorithms. Lack of labelled training data and test data may be seen as disadvantage, but in real life applications, the data is unlabelled and it's too hard to distinguish the difference between the normal and abnormal behaviour. That's why unsupervised methods are most preferred methods in real-life because they are the most proper methods to real-life applications.

CHAPTER THREE
UNSUPERVISED ANOMALY DETECTION

To be parallel to increasing importance of anomaly detection, increasing on data sizes in real time, makes difficult to distinguish the outliers or anomalies from the data. Data scientists compete with the time in analysing and cleaning of data, missing anomalies can be very harmful for a system. Unsupervised anomaly detection is the process of finding uncomfortable data points or patterns in a dataset without using any labelled and trained data. This key component of unsupervised anomaly detection methods provides real time detection on anomalies. Unsupervised anomaly detection methods are the most proper methods for real time application domains, that's why these methods are the most applicable in real life applications than the other anomaly detection approaches. But also, these methods are the most challengeable parts of anomaly detection concept. Algorithm performance has big importance at this type anomaly detection. Type of data has a key role in choosing of algorithm on the performance of algorithm.

Unsupervised anomaly detection approach is becoming more popular in recently and many different algorithms have been developed in this area. Researchers and data scientists work on more efficient unsupervised anomaly detection algorithms and improve on current algorithms. Uncontrolled anomaly detection algorithms can be classified into three categories according to the underlying theory such as nearest-neighbour based algorithms, cluster-based algorithms, statistical-based algorithms.

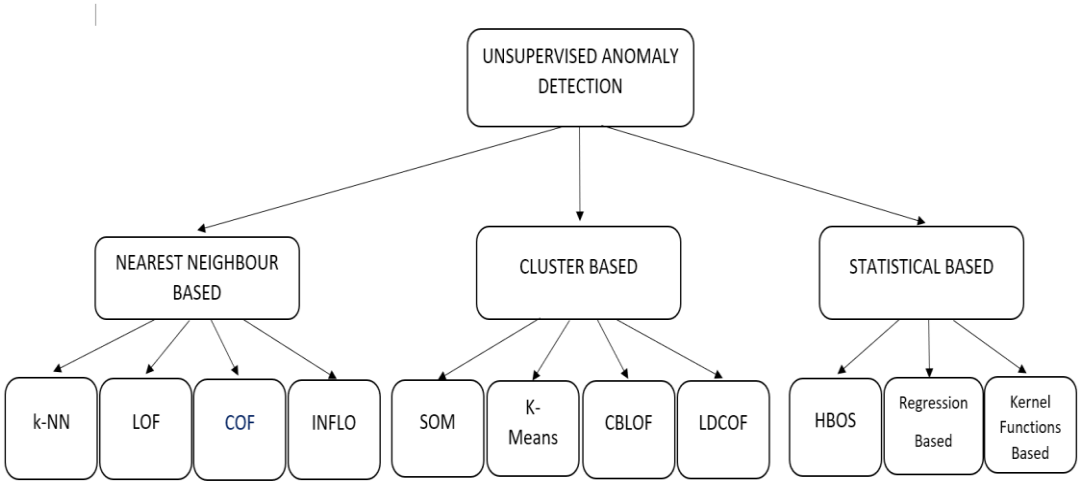


Figure 3.1 Unsupervised anomaly detection algorithms

Unsupervised anomaly detection algorithms that are shown in the Figure 3.1 is the most successful and most applicable algorithms for every approach. In this chapter, unsupervised anomaly detection approaches are given with the details.

3.1 Nearest-Neighbour Based Algorithms

Nearest-neighbour based category is the one of the important categories of unsupervised anomaly detection. Key assumption of this category is that a normal data instances lie near to the closest neighbourhood, but anomaly instances lie far from the closest neighbourhood. To determine if a data instance is anomaly or not, these algorithms use two main concepts: Distance or similarity. Distance concept is accepted more than the similarity concept in literature. Type of distance is chosen according to type of the data. For example, if data is continuous, then mostly Euclidean distance is preferred, but this does not mean that other distance measurements cannot be used. Nearest-neighbour based algorithms are in two categories as global or local. Global approaches use distance of a data instance to its k^{th} nearest neighbour, local approaches use relative density of a data instance.

Algorithms that use k^{th} distance of a data instance are the basic algorithms in Nearest-Neighbour based algorithms. These algorithms calculate anomaly score of a data instance according to its distance to its k^{th} nearest-neighbour. A threshold value

is used to determine if a data instance is anomaly or not and this threshold value is usually determined as k=1 for basic algorithms. For more complex algorithms, threshold values are determined according to different methods like Receiver Operator Characteristics curves (ROC Curves). Many different k^{th} distance-based algorithms have been developed from this basic algorithm for different data types or different anomaly types or to improve the efficiency of algorithm. Some algorithms computes number of nearest-neighbours of a data instance which is another way computing of global anomaly score. The most used global nearest-neighbour based algorithms are k-NN and k^{th}-NN (Chandola et al., 2009).

Relative density-based algorithms detect local anomalies with the density of the nearest-neighbour of every data instance. At this type algorithms don't calculate only nearest-neighbourhood of an instance, they also calculate density level of related data instance to its k^{th} neighbourhood. Density-based algorithms assume that, anomalies lie in low density of neighbourhood of data instances. Density-based algorithms also calculate anomaly score and compares this score with a threshold value. The most used and popular algorithm at this category is Local Outlier Factor (LOF) algorithm. Many other algorithms are variant of LOF algorithm. Other density-based algorithms are Connectivity Based Outlier Score (COF), Influenced Outlierness (INFLO).

Nearest-neighbour based algorithms have many advantages on anomaly detection. One of important advantage of this type algorithms is that they are naturally unsupervised, because they are data driven. Another advantage is that they can be proper to different data types, only need is to determine the right distance measure for data type. Besides advantages, they have some disadvantages to perform. For example, if a normal data instance lies far from its closest nearest-neighbour, then it can be marked as anomaly, this causes to wrong labelling normal and abnormal data instances and this can cause to missing anomalies. Also, choosing right distance measure is a key component at nearest-neighbour based algorithms. If incorrect or improper distance measure is used, then difference between normal and abnormal for dataset cannot be defined right. The best-known algorithms in this approach are k-Nearest Neighbour Based (k-NN) algorithm and Local Outlier Factor (LOF) algorithm.

k-NN algorithms is one of the global algorithms. This algorithm uses the k^{th} nearest neighbour of an instance to measure its distance to this k^{th} nearest neighbour. This algorithm is the basic nearest neighbour-based algorithm and proposed by Cover and Hart in 1967 at first time with the name of k nearest neighbour (Cover & Hart, 1967). Algorithm uses a threshold value to measure data instance's distance to its nearest neighbour. Many authors use *k=1* as threshold value to measure distance and decide about data instance according to this value if the data instance is anomaly or not. Sometimes *k* value is determined between 10<*k*<50 according to rule of thumb (Goldstein & Uchida, 2016). But choosing determination way of threshold value *k* according to structure of the data set is the truest way for anomaly detection. *k* threshold value is the key point for this algorithm. This algorithm begins with determination of *k*. Then applying the following formula, the anomaly score of a data instance is calculated (Amer, 2011):

$$knn(p) = \frac{\sum_{o \in N_k(p)} d(p,o)}{|N_k(p)|} \tag{3.1}$$

Where $knn(p)$ shows the final anomaly score, $N_k(p)$ is the k-neighbourhood and $\sum_{o \in N_k(p)} d(p,o)$ is the k-distance. Mostly Euclidean distance is used for this algorithm. Euclidean distance states difference between two points in a straight line like *p* and *o*.

Local outlier factor (LOF) algorithm is one of the very popular and most used local anomaly detection algorithms. This algorithm was proposed in 2000 by Breunig et. Al (Breunig, Kriegel, Ng, & Sander, 2000). Many researchers use LOF algorithm at different application domains to detect anomalies and consequently many variants of LOF have been proposed for anomaly detection. LOF algorithm is a nearest neighbour-based algorithm and uses k-nearest neighbour. LOF algorithm detects local anomalies, that's why there is one more step different from k-NN algorithm. LOF algorithm detects anomalies to compare local density of the data instance to its k-nearest neighbours.

To calculate LOF anomaly score, reachability distance is computed at the first step. Reachability distance is the maximum value of the between k-nearest neighbour and distance (p, o). Then taking the inverse the instance p according to $N_k(p)$, local reachability density (LRD) is computed (Goldstein & Uchida, 2016).

$$lrd_{N_k(p)}(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} reach\text{-}dist(p,o)} \tag{3.2}$$

With the computing of LRD value, LOF anomaly score that is the ratio between the related data instance and LRD value can be calculated as follows (Goldstein & Uchida, 2016):

$$LOF_{N_k(p)}(p) = \frac{\sum_{o \in N_k(p)} lrd_{N_k(o)}(o)}{|N_k(p)| * lrd_{N_k(p)}(p)} \tag{3.3}$$

Similar to k-NN algorithm, a threshold value is determined to evaluate if the data instance is anomaly or not. If the data is normal, LOF threshold value is used as 1 (Amer, 2011). If LOF score of data instance is over the 1, then the data instance is anomaly, if LOF score of data instance under the 1, then the data instance is normal. Also, some researchers use neighbourhood size range to choosing lower and upper bounds for anomaly evaluation of data instance, if data instance is in these upper and lower bounds, than this data instance is normal but if data instance is out of these upper and lower bounds, than this data instance is anomaly.

3.2 Clustering Based Algorithms

Clustering based algorithms generally accept that anomaly instances create groups. These type algorithms are unsupervised in their nature, but they can be run on semi-supervised mode at the same time. Clustering based algorithms have three assumptions and these assumptions vary according to their anomaly detection acceptance (Chandola et al., 2009).

First assumption of clustering-based algorithms accepts that normal data instances are belong to a cluster, but anomaly instances are not belonging to any cluster. This assumption is the simplest category of clustering-based algorithms, because if a data instance does not belong to any cluster, then it's an anomaly. In clustering-based anomaly detection, the key component is the clustering algorithms and many developed clustering algorithms do not need to force data instances to be in a cluster. The most used algorithms in this category are DBSCAN, ROCK, and SNN.

Second assumption of clustering-based algorithms is that normal data instances lie close to its closest cluster's centroid, while anomalies lie far from their closest cluster's centroid. This type clustering-based algorithms calculates anomaly score in two steps:

1. First data instances are separated into clusters using a proper clustering algorithm.
2. Second the distance of data instances to their cluster's centroid is calculated.

The most used algorithms in this type clustering-based anomaly detection are Self-Organizing Map (SOM) and k-Means algorithms. Algorithms in this category fail if anomalies create clusters.

Third and last assumption of clustering-based algorithms is that normal data instances are belong to large clusters, while anomalies are belonging to small clusters. In this type algorithms, first clusters are created using proper clustering algorithm; then, a threshold value is determined to mark clusters as anomaly cluster or normal clusters. If a cluster's size or density below this threshold value, then it's anomaly. CBLOF (Cluster-Based Local Outlier Factor) and LDCOF (Local Density Cluster-Based Outlier Factor) algorithms are most used algorithms of this type category. Many clustering algorithms have been developed to increase efficiency of anomaly detection performance of these type algorithms.

A similarity can be considered between nearest-neighbour based algorithms and clustering-based algorithms in terms of distance/density calculation. Both of these algorithms use distance measure, but nearest-neighbour based algorithms check the

neighbourhood of data instances, while clustering-based algorithms are checking if data instances belong to any cluster and/or its belonging level.

One of the key advantages of using clustering-based algorithms is that they can be operated in unsupervised mode naturally. Another key advantage of working with clusters is, while working with clusters, researchers only check that data instances belong to any cluster or not, or they check if data instances belong to a small or a large cluster. Working with clusters are advantage but at the same time there is a big disadvantage of clusters. If researcher does not choose a proper clustering algorithm, then anomaly detection performance will be failed dramatically. Also, because some clustering algorithm forces data instances to be in a cluster, this situation may cause to wrong marking of data instances which means, an anomaly point can be marked as a normal. There are many clustering-based algorithms and the one of the well-known clustering algorithms is Local Density Cluster-Based Outlier Factor (LDCOF) algorithm.

LDCOF algorithm is the one of the cluster-based unsupervised anomaly detection algorithms. This algorithm detects local anomalies similar to the LOF algorithm (Goldstein & Uchida, 2016) that's why this algorithm is one of the popular unsupervised anomaly detection algorithms. The main idea of the algorithm is to capture the size of the all clusters where data instances are in these clusters and the distance of data instances to their cluster centroid. Clusters need to be created before calculation of LDCOF anomaly score. LDCOF anomaly score is calculated according to cluster centroid distance of an instance, k-means clustering method is preferred to create clusters, because k-means clustering uses the assumption of clustering which accepts that the normal data instances lie close to their closest cluster centroid while the anomaly instances lie far from their closest cluster centroid. Other reasons for using of k-means clustering are easy and fast computation. Small and large clusters are created with k-means clustering at the first step and then average distance for clusters is computed to compute LDCOF anomaly score (Amer, 2011).

$$distance_{avg}(C) = \frac{\sum_{i \in C} d(i, C)}{|C|} \quad (3.4)$$

Where $|C|$ is the difference between small clusters and large clusters. After average distance is computed, LDCOF anomaly score is computed as follows:

$$LDCOF(p) = \frac{d(p, C_i)}{distance_{avg}(C_i)} \quad (3.5)$$

Where $d(p, C_i)$ is the distance of p^{th} data instance to i^{th} cluster's centroid.

LDCOF anomaly score of an instance is evaluated according to a threshold value. This threshold value is determined using rule of thumb, or if data is normal, 1 is accepted as threshold value.

3.3 Statistical Based Algorithms

Statistical-Based anomaly detection algorithms are considered under two categories: Parametric Methods and Non-Parametric Methods. But two categories have a common assumption about anomaly detection. The key assumption for statistical-based algorithms is that normal data instances are in high probability region in a probability model, while anomaly instances are in low probability region in probability model. Both parametric and non-parametric methods depend on a statistical model especially for normal behaviour of data and then they make inference about anomaly instances from the model. The difference between parametric and non-parametric methods is that parametric methods assume that they have knowledge about distribution of data and with this assumption they make inference about parameters, but non-parametric methods do not assume that they have not knowledge about distribution of data.

Parametric methods use the assumption that normal data instances come from a parametric distribution with the known parameters and known distribution as it's mentioned before. Parametric methods have sub-categories such Gaussian Model

Based, Regression Model Based, Mixture of Parametric Distribution Based. Gaussian models use maximum likelihood estimates (MLE) in basically. Regression based models depend on a regression model that model is fitted to the data and this type methods are preferred at time series data mostly. Mixture methods use mixture of parametric distributions for modelling the data.

Non-Parametric methods assume that they have no knowledge about the distribution of data. They use non-parametric statistical models to make inference about data instances if they are anomaly or not, that's why these type methods don't make estimate about model before, they learn from the data. Similar to parametric methods, non-parametric methods have sub categories: Histogram Based and Kernel Function Based. Histogram based method are the simplest category of non-parametric algorithms, they based on simple histogram creation. Kernel based methods depend on Kernel functions to estimate the density approximately.

Statistical-based anomaly detection algorithms have advantages and disadvantages. Distribution of data is both advantage and disadvantage at the same time. Data distribution is advantage because, if the distribution of data is known true, then solution is almost exact, but distribution of data is evaluated as disadvantage if the distribution of data is not true, then anomaly detection process will fail. Another advantage of statistical-based algorithms is that statistical method will be associated with a statistical hypothesis and confidence interval, these components will provide flexibility while making inference about data instances, but this is a disadvantage, because, building hypothesis testing and choosing right test statistics are not straightforward way; researcher must be very careful at this selection. The most used and the simplest algorithm in Statistical-Based algorithms is Histogram-Based Outlier Score (HBOS) algorithm.

Histograms are the simplest way to make an inference about distribution of a data in statistic. Some researchers purpose histograms as a density estimator (Scott, 1992). Creating a histogram is simple and fast, also making a comment about data from a histogram is easy for a researcher. Because of these specialties of histograms, they are

very popular on anomaly detection recently, especially they are preferred on network anomaly detection; with another word, histograms are cool guys on intrusion detection. Time is the most critical component on detection of anomalies at intrusion detection (Kind, Stoecklin, & Dimitropoulos, 2009). In this thesis, a newly proposed technique which is Histogram-Based Outlier Score (HBOS) is examined in terms of its anomaly detection performance (Goldstein & Dengel, 2012) and an adjustment is proposed to HBOS algorithm. Another main topic on histograms for anomaly detection is selection of bin-width determination technique. Selection of bin-width has big effects on anomaly detection with histogram-based techniques. Many bin-width determination techniques have been developed but the oldest bin-width determination technique was proposed by Sturges in 1926 (Sturges, 1926). Three traditional bin-width determination techniques and dynamic bin-width determination technique are examined to compare their performance on anomaly detection of histogram-based outlier score cover of this thesis.

Histogram-based anomaly detection is one of the non-parametric statistical based algorithms. A histogram provides a density estimation of univariate data. Histogram-based techniques are based on basic histogram construction and use density estimations or frequencies. These techniques can be used in both scoring-based and labelling-based. A variant of the histogram-based technique called HBOS was presented by Goldstein and Dengel in 2012 (Goldstein & Dengel, 2012). HBOS assigns anomaly score to each data instance based on the density estimation of the bin which it falls. The proposed variation of histogram-based outlier calculation HBOS algorithm can be described as follows: First histograms are constructed for each variable. Then, dynamic bin-width is determined as follows: A single bin is determined, and fixed amount of successful values are stored in this bin. Fixed bin is computed as $\frac{N}{k}$, where N is sample size and k is the number of bins. The number of bins is same here, because the number of bins is determined at first, the height of bins is computed. The large interval for values means that low density for histogram height. Histograms are calculated for every variable and then normalization process is applied to histograms to bring height of every histogram to 1 and this normalization process

provides equal contribution of data instances to anomaly score. Finally, anomaly score of every instance is calculated as follows (Goldstein & Dengel, 2012):

$$HBOS(p) = \sum_{i=1}^{d} \log(\frac{1}{hist_i(p)}) \tag{3.6}$$

Where, *d* is the number of variables, *p* is the p^{th} data instance and $hist_i(p)$ is the standardized density estimation of the corresponding bin which p^{th} data instance falls.

3.4 Adjusted Histogram Based Outlier Score (Adjusted HBOS)

The critical point in HBOS algorithm is the bin-width used when constructing the histogram. Small bin-widths may increase the false positive ratio by leading to the evaluation of normal instances as anomalies. If the bins are too wide, the false negative ratio may increase. Because the anomalies may be evaluated as normal instances. Therefore, the bin-width rule can greatly affect the success of the anomaly detection algorithm. Dynamic bin-width is the one of the state of art bin-width determination technique recently. But also, still traditional techniques are valid and usable for bin-width determination. The most used traditional bin-width determination techniques are Sturges Rule, Scott's Rule, Freedman and Diaconis (FD) Rule.

In dynamic bin-width determination, technique uses the number of bins, *k.* The data points are discretized into a number of bins, and the number of points in each bin is used to compute outlier score.

Sturges rule was proposed in 1926 and the oldest method at histogram bin-width determination (Sturges, 1926). This technique was derived from a binomial distribution and assumes that this binomial distribution is approximately normal. Sturges rule formula as follows (Sturges, 1926):

$$k = [log_2 n] + 1 \tag{3.7}$$

Where *k* is the bin-width and *n* is the number of instances.

Second most used traditional bin-width determination technique is Scott's rule. It was proposed in 1979 (Scott, 1979). This technique uses standard deviation of data samples and assumes that these data samples are normally distributed. Scott's rule formula as follows (Scott, 1979):

$$k = \frac{3.5\hat{\sigma}}{\sqrt[3]{n}} \tag{3.8}$$

Where *k* is bin-width, *n* is the number of instances and $\hat{\sigma}$ is standard deviation of samples.

The another most used traditional bin-width determination technique is Freedman and Diaconis rule. This technique was proposed in 1981 (Freedman & Diaconis, 1981). Technique uses interquartile range of the data (IQR). FD rule as follows (Freedman & Diaconis, 1981):

$$k = 2\frac{IQR(x)}{\sqrt[3]{n}} \tag{3.9}$$

Where *k* is bin-width, *n* is the number of instances and *IQR(x)* is the interquartile range of the data.

An adjustment is proposed in HBOS algorithm for the situations where min-max normalization is used. As it's mentioned before, proposed HBOS formula is computed like below:

$$HBOS(p) = \sum_{i=1}^{d} \log(\frac{1}{hist_i(p)}) \tag{3.10}$$

Min-max normalization brings the all attributes into [0,1] range and in this formula $hist_i(p)$ is the standardized density estimation of the corresponding bin which p^{th} data instance falls. When we use min-max normalization we may get division by zero

uncertainty. In this case, Goldstein and Dengel use 0 value on their algorithm calculations (Goldstein & Dengel, 2012). But assigning 0 value causes to loose contribution of data instance to the anomaly score. As a result of losing correct contribution to the anomaly score, missing anomalies or wrong marking of data instances may be occurred.

We proposed an adjustment to prevent division by zero uncertainty when using min-max normalization. Proposed adjustment depends on the sample size. The adjusted HBOS is given below:

$$HBOS(p) = \sum_{i=1}^{d} \log\left(\frac{1}{hist_i(p) + \frac{1}{n}}\right) \tag{3.11}$$

The aim of this adjustment is to prevent division by zero uncertainty in the standard HBOS. This adjustment also prevents wrong evaluation and wrong classification of a data instance as an anomaly.

CHAPTER FOUR
APPLICATION

This section covers two applications. In the first application, four unsupervised anomaly detection algorithms are compared in terms of accuracy and sensitivity. In the second application, HBOS algorithm is considered in detail and histogram bin-width determination techniques are compared in terms of their effects on the performance of HBOS algorithm. Datasets are chosen from both synthetic and real datasets that are published publicly. Only point anomalies are considered. RapidMiner is used for the first application. R statistical programming language is used to implement HBOS algorithm with various bin-width determination techniques.

4.1 Datasets

Sixteen datasets are used in total for two applications. Two of datasets which are Wisconsin Breast Cancer and KDD99, are joint for two applications, but modified versions of these two datasets are used for second application. Two of datasets which are Shuttle and KDD99 are synthetic datasets, they are produced artificially for supervised anomaly detection tasks, but they are mostly used for benchmarking unsupervised anomaly detection algorithms performance. Other fourteen datasets are real value datasets that are collection of real-life observations. All datasets except the Yahoo A1 Benchmark datasets are multivariate. Detailed information about the datasets are given below in Table 4.1 and Table 4.2.

Table 4.1 The datasets used for benchmarking the unsupervised anomaly detection algorithms. Datasets are publicly published at related sources. They all have different outlier (anomaly) percentages

Dataset	Sample Size	Outlier Percentage	Source
Wisconsin Breast Cancer	569	0.37%	UCI
Real_32	1427	0.03%	Yahoo A1 Benchmark
Real_28	1441	0.06%	Yahoo A1 Benchmark
Real_46	1441	0.08%	Yahoo A1 Benchmark
Arrhythmia_231	1571	0.81%	PhysioBank Databases
Arrhythmia_106	2027	0.47%	PhysioBank Databases
Arrhythmia_220	2048	0.08%	PhysioBank Databases
KDDCUP99-10% Corrected	494021	0.80%	UCI

Table 4.2 The list of the datasets for the second application. The datasets are modified by Goldstein for benchmarking unsupervised anomaly detection methods (Goldstein, 2015). They all have different outlier percentage

Dataset	Sample Size	Outlier Percentage	Source
Breast Cancer	367	2.72%	http://dx.doi.org/10.7910/DVN/OPQMVF
Pen-Global	809	11.1%	
Letter	1,600	6.25%	
Speech	3,686	1.65%	
Satellite	5,100	1.49%	
Pen-Local	6,724	0.15%	
Annthyroid	6,916	3.61%	
Shuttle	46,464	1.89%	
Aloi	50,000	3.02%	
KDDCUP99	620,098	0.17%	

Wisconsin Breast Cancer (WBC) dataset is one of the popular unsupervised anomaly detection datasets. It consists of real observations, there are 32 attributes in

the dataset. The dataset is used to detect anomalous cancer cells from real patient values, all features that are used for evaluation are obtained from digital images.

Pen Based Recognition of Handwritten Text_Global dataset consists of real observations from 44 different writers. The dataset has 16 attributes. In the modified global version, 8 of digits are accepted as normal and 10 digits are accepted as anomaly.

Yahoo A1 Benchmark datasets (Real_32, Real_28, real_46) are time series. They consist of real observations.

Letter Recognition dataset is a real observation dataset which has 16 attributes. Researchers targeted to identify 26 capital letters in English alphabet as black and white rectangular pixels.

Arrhythmia datasets are composed of electrocardiogram observations from Massachusetts Institute of Technology Beth Israel Hospital. Arrhythmia is a problem of heart beating rhythm. The datasets are gathered from 47 patients in the hospital between 1975 and 1979. The datasets have 5 attributes.

Speech Accent dataset consists real observations and it's about native and non-native English speakers. The observations belong to non-native English speakers are accepted as anomaly. Dataset has 400 attributes.

Landsat Satellite dataset is composed of real satellite observations. It has 36 attributes.

Pen Based Recognition of Handwritten Text_Local dataset consists of real observations. The dataset has 16 attributes. The data is collected from 44 different writers. In this dataset, all digits except 4 digits are kept unlike the global dataset.

Thyroid Disease dataset is another real dataset. the observations are collected from thyroid patients. It has 21 attributes.

Statlog Shuttle dataset is one of two artificial produced datasets generated for supervised anomaly detection task but as it's mentioned before, it's used for benchmarking unsupervised anomaly detection methods. It has 9 attributes, and it has been produced by NASA.

Object Images_Aloi dataset is gathered from Amsterdam Library of Object Images. It has 27 attributes.

KDDCUP99 (Kdd99) is one of the artificial produced datasets and this dataset is also mostly used in intrusion detection. This dataset has been produced for supervised anomaly detection tasks, but it can be used for benchmarking unsupervised anomaly detection methods. It has 38 attributes.

4.2 Comparison of Unsupervised Anomaly Detection Algorithms

Aim of the first application is to compare the performances of four selected unsupervised anomaly detection algorithms in terms of accuracy and sensitivity (Kizilkaya & Yildiztepe, 2018). Eight datasets are used in the evaluation of performances of algorithms. In the first application RapidMiner Studio is used (RapidMiner, 2018). RapidMiner is a user-friendly software which has been developed for data analysis and data mining. It has an unsupervised anomaly detection extension. The extension includes sixteen anomaly detection algorithms.

4.2.1 Application Design

In this application four unsupervised anomaly detection algorithms are chosen for benchmarking. Two of these algorithms (k-NN and LOF) are in nearest-neighbour based category. The third algorithm LDCOF is a cluster-based algorithm. The latter

one (HBOS) is a statistical-based algorithm. Only k-NN is a global anomaly detection algorithm among these algorithms.

The details of the RapidMiner application for the first data set are given below in Figure 4.1. The same process is repeated for other data sets and these process are shown in Figure 4.2, 4.3 and 4.4. The initial parameters of each algorithm must be set. The selected values of the parameters can be seen in the following ways. The mixed Euclidean measure is used for distance-based algorithms. Because the datasets include both categorical and numerical variables.

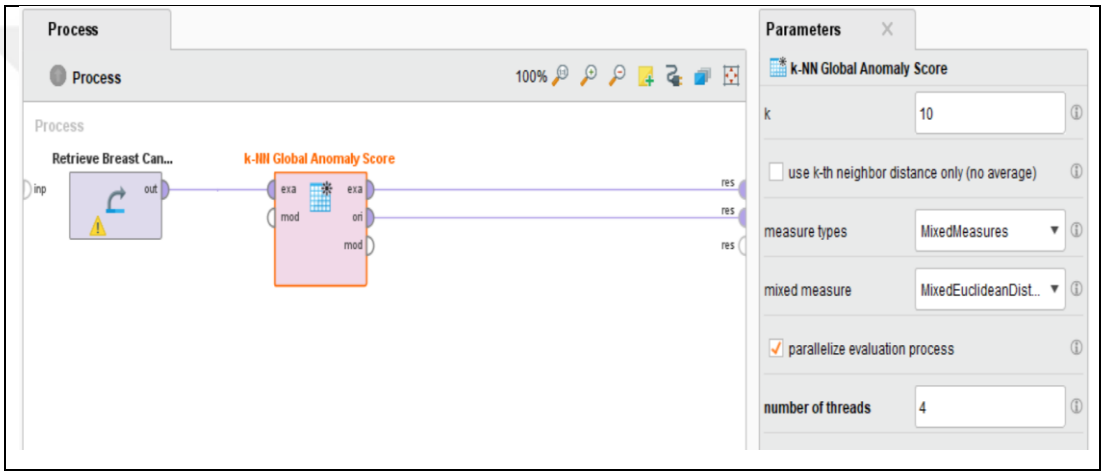


Figure 4.1 k-NN algorithm design for the WBC dataset in RapidMiner

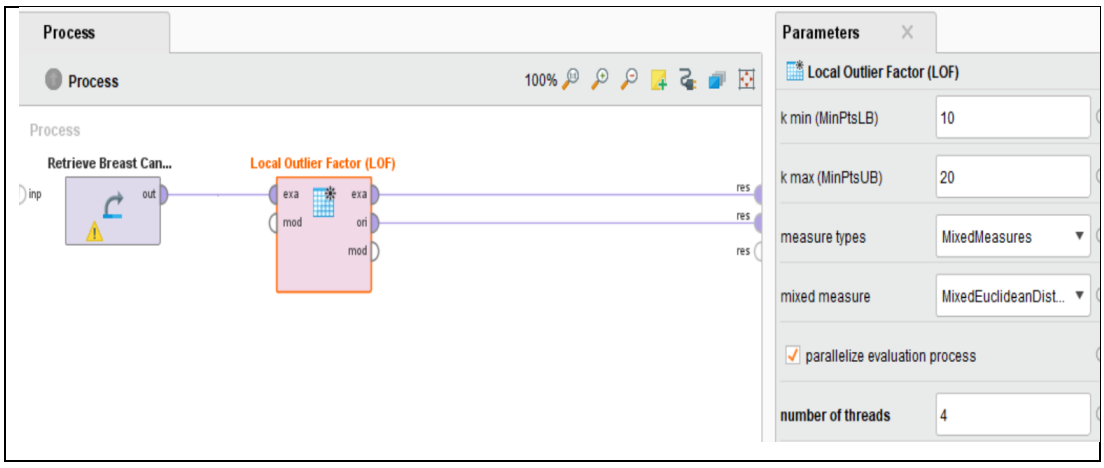


Figure 4.2 LOF algorithm design for the WBC dataset in RapidMiner

Because LOF algorithm is a local anomaly-based algorithm, an interval is determined for nearest-neighbourhood with using the rule of thumb and Mixed Euclidean measure is used.

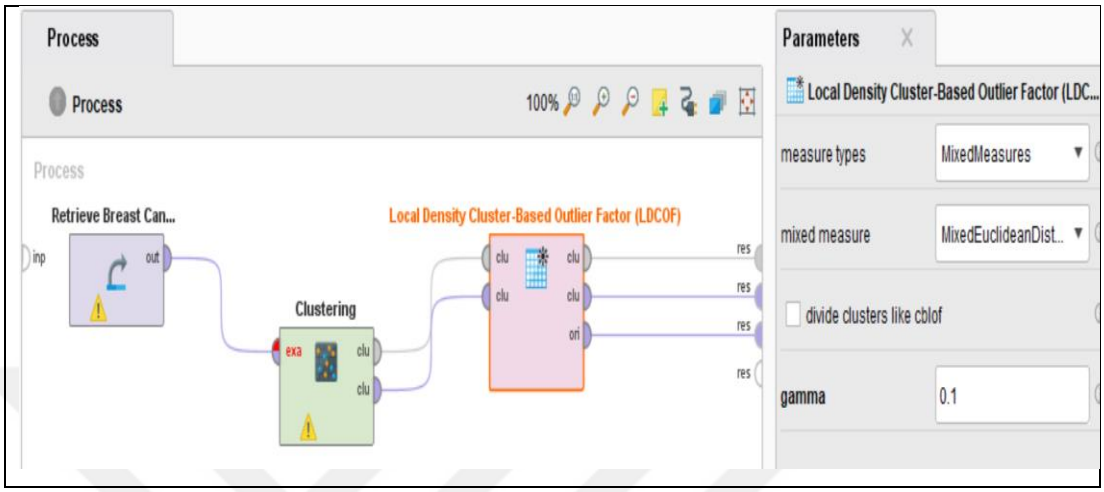


Figure 4.3 LDCOF algorithm design for the WBC dataset in RapidMiner

k-means clustering is used in the clustering step of LDCOF algorithm. Using another clustering method at this step may change the results.

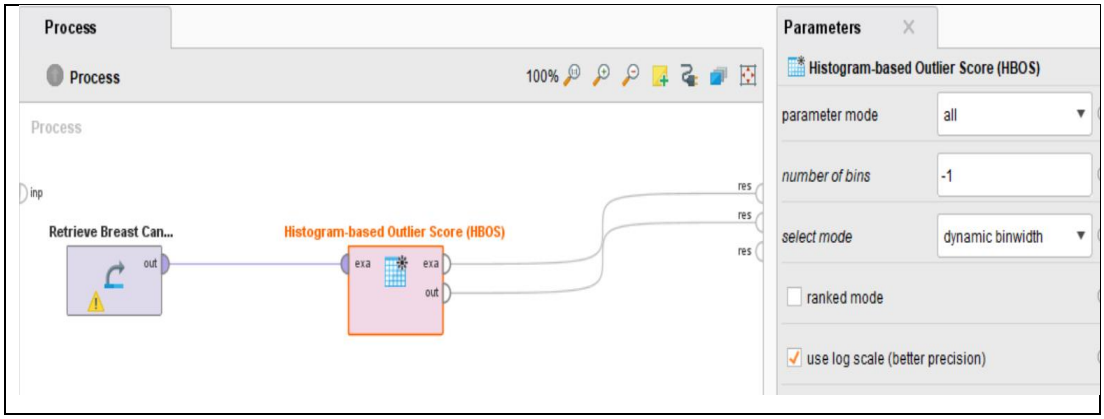


Figure 4.4 HBOS algorithm design for the WBC dataset in RapidMiner

The dynamic bin-width method proposed by Goldstein and Dengel is used in the HBOS algorithm (Goldstein & Dengel, 2012).

Anomaly scores are computed for every data instance and a score column is added to the dataset. The result can be sorted in descending order by the anomaly score for easy ranking. Figure 4.5 is an example for this descending order.

Row No.	score ↓	att1	att2	att3	att4	att5	att6	att7
109	44.940	86355	M	22.270	19.670	152.800	1509	0.133
123	43.629	865423	M	24.250	20.200	166.200	1761	0.145
79	41.341	8610862	M	20.180	23.970	143.700	1245	0.129
462	39.331	911296202	M	27.420	26.270	186.900	2501	0.108
568	39.295	927241	M	20.600	29.330	140.100	1265	0.118
1	37.508	842302	M	17.990	10.380	122.800	1001	0.118
83	36.143	8611555	M	25.220	24.910	171.500	1878	0.106
213	34.637	8810703	M	28.110	18.470	188.500	2499	0.114
4	34.565	84348301	M	11.420	20.380	77.580	386.100	0.142
43	34.280	855625	M	19.070	24.810	128.300	1104	0.091
13	34.156	846226	M	19.170	24.800	132.400	1123	0.097
259	34.039	887181	M	15.660	23.200	110.200	773.500	0.111
153	32.384	8710441	B	9.731	15.340	63.780	300.200	0.107
564	31.820	926125	M	20.920	25.090	143	1347	0.110

Figure 4.5 Dataset view after the anomaly score computation

4.2.2 Results of the First Application

The accuracy and sensitivity results of the algorithms are given in this section. Accuracy is a performance metric and high accuracy indicates high precision but not enough to evaluate a method alone. Accuracy is calculated with the ratio of the total number of true positives and true negatives, to the sample size. Sensitivity is another metric for performance measurements that indicates the ability to determine anomaly points correctly. It is also called the true positive rate. Sensitivity is calculated with the ratio of true positives to the total number of true positives and false negatives. (Fawcett, 2005). The results are given in Table 4.3 and Table 4.4.

Table 4.3 Accuracy rates

Dataset	**k-NN**	**LOF**	**LDCOF**	**HBOS**
WBCD	0.6161	0.4446	0.3732	0.3732
KDDCUP99	0.1593	0.8160	0.1600	0.1761
Yahoo Real_28	0.0388	0.1887	0.6217	0.9743
Yahoo Real_32	0.9600	0.0328	0.5017	0.5998
Yahoo Real_46	0.0117	0.1963	0.5648	0.9673
Arrhythmia_106	0.5891	0.4583	0.5934	0.5678
Arrhythmia_220	0.9248	0.2016	0.5380	0.9648
Arrhytyhmia_231	0.1960	0.7313	0.5556	0.1425

Table 4.4 Sensitivity rates

Dataset	**k-NN**	**LOF**	**LDCOF**	**HBOS**
WBCD	0.2028	0.9622	0.1082	0.9952
KDDCUP99	0.0007	0.9832	0.0061	0.0091
Yahoo Real_28	0.6829	0.8902	0.9756	1.0000
Yahoo Real_32	0.2765	0.9574	0.5106	1.0000
Yahoo Real_46	0.1376	0.8532	0.9724	1.0000
Arrhythmia_106	0.2207	0.8532	0.4470	0.1186
Arrhythmia_220	0.0064	0.8580	0.1483	0.8516
Arrhytyhmia_231	0.0055	0.8669	0.5251	0.0984

According to the accuracy rate for WBC dataset, k-NN algorithm is the most successful algorithm to classify anomaly instances and LOF algorithm has the second successful performance. LDCOF and HBOS algorithms have lowest accuracy rates for the WBC dataset. But in terms of sensitivity rates, HBOS algorithm is the most successful algorithm for WBC dataset and k-NN algorithm has very low sensitivity rate for this dataset. LDCOF algorithm has the lowest sensitivity rate for WBC dataset.

For Kdd99 dataset, LOF algorithm is the most successful algorithm in terms of both accuracy and sensitivity rate. The other algorithms have shown very low accuracy and sensitivity rates.

For Yahoo Real_28 dataset, HBOS algorithm is the most successful algorithm. According to accuracy rate and sensitivity rate, LDCOF is the second successful algorithm after HBOS. k-NN algorithm has the lowest rates for Yahoo Real_28 dataset in terms of both accuracy and sensitivity.

In terms of accuracy, the k-NN algorithm has shown the most successful performance for Yahoo Real_32 dataset, but in terms of sensitivity, the most successful algorithm is HBOS for Yahoo Real_32 dataset. LOF algorithm has the lowest success according to accuracy while the k-NN algorithm has the lowest success in terms of sensitivity for this dataset.

Similar to Yahoo Real_28 dataset results, the most successful algorithm in terms of both accuracy and sensitivity is HBOS for Yahoo Real_46. k-NN algorithm give the worst performance for this dataset in terms of both accuracy and sensitivity.

For Arrhythmia datasets (Arrhythmia_106, Arrhythmia_220 and Arrhytyhmia_231), LDCOF, HBOS and LOF give successful accuracy values respectively. But in terms of sensitivity, LOF shows the best performance.

In general, according to the results obtained in this section, it can be said that HBOS and LOF algorithms perform better than two other algorithms in terms of accuracy and sensitivity. To decide whether data instances are anomaly or not, default RapidMiner threshold values are used. Also, default parameter values are used for the algorithms. Choosing the ideal parameter value is a critical point and effects the results. But tuning the parameters increases the computation time.

4.3 Application of HBOS

HBOS is one of the fast-unsupervised algorithms. It can be applied to univariate or multivariate data. There are two key components that effect the performance of HBOS. The most important component is the bin-width determination techniques used to specify bins. The latter component is the normalization method which should be applied to the variables before constructing histograms. In this application, the effect of the various bin width determination techniques to the performance of the HBOS is investigated (Kizilkaya & Yildiztepe, 2018).

4.3.1 Application Design

In this application, four bin-width determination techniques (Sturges, Scott's, F-D, and dynamic bin width) are compared in terms of contribution to the performance of the adjusted HBOS algorithm. All density values computed by histograms are normalized using min-max method to provide equal contribution of data instances to anomaly score. The ten datasets which are modified by Goldstein (Goldstein, 2015) for unsupervised anomaly detection tasks are used in the study and these datasets can be seen in Table 4.5. The detailed information about the datasets is given in section 4.1.

Table 4.5 The ten datasets used for the second application. The datasets are modified by Goldstein for benchmarking unsupervised anomaly detection algorithms (Goldstein, 2015). They all have different outlier percentages

Dataset	**Sample size**	**Number of Variables**	**Number of Outliers**	**Outlier percentage**
Breast cancer	367	30	10	2.72
Pen-global	809	16	90	11.1
Letter	1,600	32	100	6.25
Speech	3,686	400	61	1.65
Satellite	5,100	36	75	1.49
Pen-local	6,724	16	10	0.15
Annthyroid	6,916	21	250	3.61
Shuttle	46,464	9	878	1.89
Aloi	50,000	27	1508	3.02
Kdd99	620,098	38	1,052	0.17

4.3.2 Results of the Second Application

HBOS algorithm computes an anomaly score for every data instance. A threshold value must be set to assign a label to the data instances as an anomaly. In this study, the threshold values are obtained by using receiver operator characteristic (ROC) curves. These threshold values are those that maximize the area under the curve (AUC). Table 4.6 shows the threshold values for every bin-width determination technique and dataset. Data instances with score values exceeding the threshold are accepted as an anomaly. Two examples of the ROC curves are given in Figure 4.6 and Figure 4.7.

Table 4.6 Threshold values for the anomaly scores

Dataset	**Sturges**	**Scott's**	**F-D**	**Dynamic bin width**
Breast cancer	16.5	20	21	19
Pen-global	8	8.5	8.5	10
Letter	10.5	12	12	11.5
Speech	69	77	77	85.5
Satellite	14	17	16.5	18.25
Pen-local	9.25	10.18	9.99	15.62
Annthyroid	4.3	6.5	6.5	6.5
Shuttle	4	8	8	5.5
Aloi	1.5	5	5	5
Kdd99	10	16.5	19	18

.

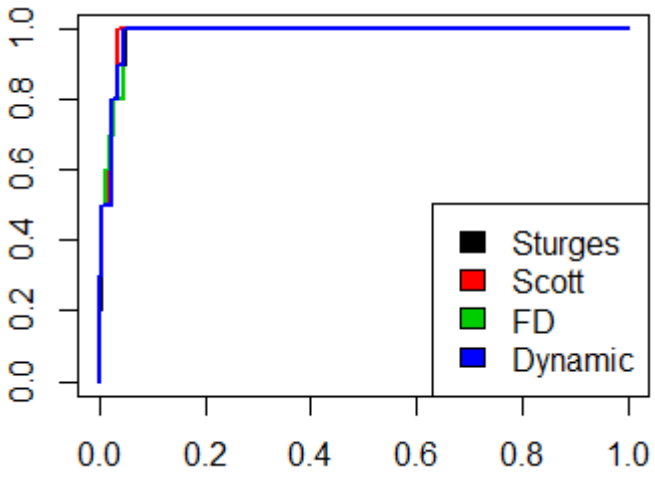


Figure 4.6 ROC curve for the WBC dataset

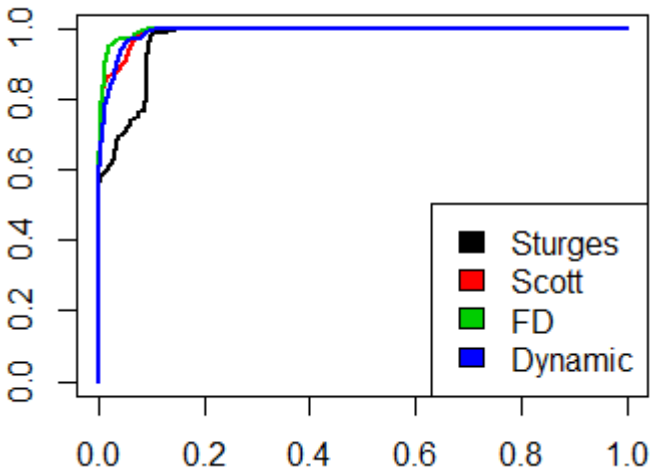


Figure 4.7 ROC curve for the KDD99 dataset

Table 4.7 shows the performance measures of the best resulted bin-width determination technique for every dataset. True positive rates (TPR) is the proportion of actual anomalies which are correctly detected as anomaly. False positive rates (FPR) indicates the proportion of normal instances which are incorrectly identified as anomaly. It can be seen that Scott's rule and F-D rule are the best techniques for four datasets. Dynamic bin-width performs better than the other techniques only on the Annthyroid dataset based on accuracy rate.

Table 4.7 Performance measures for the best resulted bin-width determination techniques

Dataset	**Accuracy Rate**	**TPR**	**FPR**	**Best Technique**
Breast cancer	0.9618	1.0000	0.0392	Scott's
Pen-global	0.7119	0.6222	0.2767	F-D
Letter	0.6331	0.5800	0.3630	F-D
Speech	0.2563	0.7704	0.7522	F-D & Scott's
Satellite	0.9505	0.7600	0.0465	Scott's
Pen-local	0.6039	0.8000	0.3963	Sturges
Annthyroid	0.7241	0.5480	0.2692	Dynamic
Shuttle	0.9931	0.9794	0.0066	Scott's
Aloi	0.4510	0.6213	0.5542	Sturges
Kdd99	0.9778	0.9534	0.0221	F-D

AUC is the integral of the ROC also can be used as a performance criterion. Table 4.8 shows the best AUC values for datasets according to used bin-width determination techniques. AUC results show that traditional techniques are more successful than the dynamic bin-width. Scott's Rule comes forward as the most successful bin-width determination technique in four compared techniques for our study for 5 datasets of 10 datasets. F-D rule and Sturges rule are the second successful bin-width determination techniques after Scott's rule. Scott's rule and F-D rule give the same AUC value for Satellite dataset. But the accuracy rate of the Scott's rule is higher. Dynamic bin width technique does not give the most successful results, but the values are very close. Also, it can be seen that there are no differences between synthetic and real data sets for bin-width determination techniques. Another result is that the sample size has no effect on the success of the HBOS. Dynamic bin-width results are compatible with those found by Goldstein and Uchida. (Goldstein & Uchida, 2016)

Table 4.8 The AUC results for the bin-width determination techniques

Dataset	**Sturges**	**Scott's**	**F-D**	**Dynamic bin width**
Aloi	0.5343	0.5048	0.4832	0.5000
Annthyroid	0.5206	0.6737	0.6616	0.6708
Breast Cancer	0.9843	0.9860	0.9843	0.9843
KDD99	0.9704	0.9909	0.9946	0.9900
Letter	0.6405	0.6331	0.6352	0.6332
Pen-Global	0.7202	0.7238	0.7245	0.7127
Pen-Local	0.7094	0.7781	0.7681	0.6939
Satellite	0.9152	0.9233	0.9233	0.9224
Shuttle	0.9809	0.9978	0.9968	0.9942
Speech	0.4701	0.4699	0.4699	0.4686

CHAPTER FIVE
CONCLUSION

In this thesis, unsupervised anomaly detection techniques are studied. The four different algorithms chosen are compared using data sets with different characteristics. The HBOS algorithm was also considered, and a correction based on the sample size is proposed as a solution to the problem of uncertainty when min-max normalization is performed in the HBOS algorithm. The effects of different histogram bin width determination techniques on the performance of the algorithm are investigated using the proposed calculation. The results are compared using various performance measures.

In the study, firstly, four unsupervised algorithms, k-NN, LOF, LDCOF, and HBOS is included in the study. In the first application, seven real datasets and one artificial dataset are used. Results are obtained by using RapidMiner Studio for the first application. Accuracy rate and sensitivity rate are used as the performance measurement of the algorithms for the datasets. There is no algorithm that performs best in all datasets since the results can be affected by the characteristics of data or selected parameters. LOF gives the best performance on the largest (Kdd99) dataset where the other algorithms obtain the poor results. For the smallest (WBC) dataset all algorithms perform average. HBOS gives the best results in terms of both accuracy and sensitivity rates for Yahoo A1 Benchmark datasets. Generally, it can be said that HBOS and LOF algorithms perform better in terms of accuracy and sensitivity.

The second application focus on the HBOS algorithm. In the second application, the effect of the various bin width determination techniques to the performance of the adjusted HBOS is examined. Performance of bin-width determination techniques are measured by using accuracy rate and AUC. Threshold values are determined for every dataset and for every technique by using ROC Curves. Ten different datasets from different application domains are used for measuring performance of HBOS. Min-max normalization is applied to the density values computed by histograms. Four techniques, which are Sturges rule, Scott’s rule and Freedman-Diaconis rule, and the

dynamic bin-width determination are evaluated. The results show that dynamic bin-width doesn’t perform better. Scott's rule is the most successful technique according to the AUC results among traditional techniques.

As a general summary, k-NN based methods and statistical base methods, in particular LOF and HBOS are recommended for unsupervised anomaly detection tasks. If the computation time is a concern HBOS may be preferred. Also, HBOS can be easily applied to multivariate data. If the density values are to be normalized with min-max, it is strongly recommended to use proposed adjusted HBOS. The applications for HBOS have also revealed the importance of the selected box width determination method. Improvement of the proposed adjustment, examine it for mixed-attribute data sets and researching new approaches about bin-width determination techniques might be the main motivation of the future studies.

REFERENCES

Ahmed, M. (2018). Collective anomaly detection techniques for network traffic analysis. *Annals of Data Science, 5*(4), 497-512.

Alestra, S., Bordry, C., Brand, C., Burnaev, E., Erofeev, P., Papanov, A., & Silveira-Freixo, C. (2014). Application of rare event anticipation techniques to aircraft helath management. *Advanced Materials Research, 1016,* 413-417.

Amer, M. (2011). *Comparison of Unsupervised Anomaly Detection Techniques.* Bachelor Thesis, German University in Cairo, Germany.

Bontemps, L., McDermott, J., & Le-Khac, N. (2016). Collective anomaly detection based on long short-term memory recurrent neural networks. *International Conference on Future Data and Security Engineering*, 141-152.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data,* 93-104.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Comput. Surveys (CSUR), 41* (3), 15.

Chuah, M., & Fu, F. (2007). ECG anomaly detection via time series analysis. *International Symposium on Parallel and Distributed Processing and Applications,* 123-135. Berlin: Springer.

Clausi, D. (2002). K-means Iterative Fisher (KIF) Unsupervised Clustering Algorithm Applied to Image Texture Segmentation. *Pattern Recognition, 35* (9), 1959-1972.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13* (1), 21-27.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *In Applications of data mining in computer security*, 77-101. Boston: Springer.

Fawcett, T. (2005). An Introduction to ROC Analysis. *Pattern Recognition Letters, 27* (8), 861-874.

Freedman, D., & Diaconis, P. (1981). On the Histogram as a Density Estimator: L2 theory. *Probability Theory and Related Fields, 57* (4), 453-476.

Gama, F., Arruda, H., Carvalho, H., Souza, P., & Pessin, G. (2017). Improving our understanding of the Behaviour of Bees Through Anomaly Detection Techniques. *International Conference on Artificial Neural Networks*, 520-527.

Gogoi, P., Borah, B., & Bhattacharya, D. (2010). Anomaly detection analysis of intrusion data using supervised & unsupervised approach. *Journal of Converge Information Technology, 5* (1), 95-110.

Goldstein, M. (2015). Unsupervised Anomaly Detection Benchmark. Retrieved December 4, 2018 from Harvard Dataverse. doi:10.7910/DVN/OPQMVF

Goldstein, M., & Dengel, A. (2012). Histogram-based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algortihm. *S. Wölfl (ED.): Poster and Demo Track of the 35th German Conference on Artificial Intelleigence (KI-2012*), 59-63. Kaiserslautern.

Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE, 11* (4), 1-31. https://doi:10.1371/journal.pone.0152173.

Golmohammadi, K., & Zaiane, O. (2015). Time series contextual anomaly detection for detecting market manipulation in stock market. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10. Paris: IEEE.

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised Anomaly Detection. In *Proceedings of the 20th international joint conference on Artificial intelligence*, 1624-1628.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning-Data Mining, Inference, and Prediction.* Stanford: Springer.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review, 22*(2), 85-126.

Jain, A. K. (2009). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters, 31* (8), 651-666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Application in R.* New York: Springer.

Kind, A., Stoecklin, M., & Dimitropoulos, X. (2009). Histogram-based Traffic Anomaly Detection. *IEEE Transactions on Network and Service Management, 6* (2), 110-121.

Kizilkaya, B., & Yildiztepe, E. (2018, October 6). Anomaly Detection in Multivariate Data Using Histogram Based Outlier Score. *11th Internaitonal Statistics Days Conference* . Muğla.

Kizilkaya, B., & Yildiztepe, E. (2018, April 28). Evaluation of Performances of Unsupervised Anomaly Detection Algorithms. *4th International Researchers, Statisticians and Young Statisticians Congress.* İzmir.

Leung, K., & Leckie, C. (2005). Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eight Australasian conference on computer science 38*, 333-342. Australian Computer Society, Australia.

Nguyen, M., Omiecinski, E., & Mark, L. (2010). *A Fast Randomized Method for Local Density-based Outlier Detection in High Dimensional Data.* Atlanta: College of Computing, Georgia Institute of Technology.

Oh, C., & Shon, H. (2009). Damage diagnosis under environmental and operational variations using unsupervised support vector machine. *Journal of sound and vibration*, *325* (1-2), 224-239.

Pawar, A. M., & Mahindrakar, M. (2015). A Comprehensive Survey on Online Anomaly Detection. *International Journal of Computer Applications*, 41-45.

RapidMiner, (n.d.). Retrieved 3 December 2018 from https://rapidminer.com/products/studio/.

Schwabacher, M., Oza, N., & Matthews, B. (2009). Unsupervised anomaly detection for liquid-fueled rocket propulsion health monitoring. *Journal of Aerospace Computing, Information and Communication*, *6* (7), 464-482.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66* (3), 605-610.

Scott, D. W. (1992). Histograms: Theory and Practice. In D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*, 47-94. John Wiley & Sons, Inc.

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, *21* (153), 65-66.

Tsai, C., Hsu, Y., Lin, C., & Lin, W. (2009). Intrusion detection by machine learning: A review. *Expert Systems with Applications, 36* (10),11994-12000.

Zhongguo, Y., Hongqi, L., Ali, S., & Yile, A. (2017). Choosing Classification Algorithms and Its Optimum Parameters based on Data Set Characteristics. *Journal of Computers, 5* (2017), 26-38.