AN ADAPTIVE OUTLIER DETECTION FOR SCATTER POINTS OF UNASCERTAINED MODELS

DAVINNA JEREMIAH

UNIVERSITI SAINS MALAYSIA

2016

AN ADAPTIVE OUTLIER DETECTION FOR SCATTER POINTS OF UNASCERTAINED MODELS

by

DAVINNA JEREMIAH

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

February 2016

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude and thanksgiving to God for without Him, the completion of this thesis would not have been possible at all. I thank Him for His faithfulness in seeing me through this journey, during which I had witnessed His greatness and mercies time and again. He indeed has been my ever present help in time of need. To Him alone be all the glory.

I would like to express my appreciation to my supervisors, Associate Professor Ong Hong Choon and Professor Low Heng Chin for their advice, time and encouragement. I thank them for all their support and for being truly understanding and patient despite my many shortcomings.

I am truly grateful to my husband for his loving support, for who has always been steadfast in ensuring the little details in life are attended to and for unfailingly, having them all done in good cheer!

Last but not least, I would like to express my heartfelt gratitude to my dear parents. If it were not for their unceasing prayers, I would not have been enabled to have kept going day after day. Their unwavering support, encouragement and love, I deeply treasure and forever I would be thankful.

TABLE OF CONTENTS

Page

CHAPTER 1 – INTRODUCTION

CHAPTER 2 – LITERATURE REVIEW

CHAPTER 3 – THE PROPOSED OUTLIER DETECTION METHOD

CHAPTER 4 – THE ANALYSIS OF THE PROPOSED CLUSTER DENSITY COMPUTATION

CHAPTER 5 – PERFORMANCE EVALUATION, RESULTS AND DISCUSSION

CHAPTER 6 – CONCLUSION

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBLOF Cluster-Based Local Outlier Factor

CMD Compact Matrix Decomposition

COF Connectivity-based Outlier Factor

EM Expectation-Maximization

FAST-MCD FAST-Minimum Covariance Determinant

GLCM Grey Level Co-occurrence Matrices

KSD Kernelized Spatial Depth

LOF Local Outlier Factor

MCD Minimum Covariance Determinant

MCMC Markov Chain Monte Carlo

PCA Principal Component Analysis

pdf probability density function

SUGS Sequential Updating Greedy Search

SVMs Support Vector Machines

LIST OF SYMBOLS

The following is a list of symbols related to the proposed method:

q index value of a data point

k index value of a cluster

K total number of clusters

z cluster assignment parameter

$z_q = k$ assignment of the q^{th} data point to cluster k

$K+1$ new cluster

α control parameter

$\boldsymbol{\theta}_k$ distribution parameter of cluster k

$\mathbf{x}_q$ a p-dimensional data point, $\mathbf{x}_q = (x_1^q, \ldots, x_p^q)$

$\mathbf{X}^{(q)}$ the set of data points until the current point, $\mathbf{x}_q$ thus, $\mathbf{X}^{(q)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$

$Z^{(q-1)}$ the set of cluster assignments until z_{q-1} thus, $Z^{(q-1)} = \{z_1, \ldots, z_{q-1}\}$

p number of dimensions

N total number of data points on the scatter plot

t threshold

w percentage of the remaining number of data points on the scatter plot

$\neg N$ total remaining data points on the scatter plot that were not eliminated

$\boldsymbol{\mu}$ a p-dimensional mean of a data cluster

$\boldsymbol{\Sigma}$ a $p \times p$ covariance matrix

$\mathbf{m}$ a p-dimensional mean of $\boldsymbol{\mu}$

d degrees of freedom

r cluster separability parameter

$\boldsymbol{\Lambda}$ a $p \times p$ precision matrix

$\mathbf{S}$ a $p \times p$ scale matrix

$\mathbf{X}$ a set of points from a cluster, $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$

n number of data points in a cluster

$\tilde{\mathbf{x}}_q$ a data point of a cluster centred at 0

$(\lambda_1\mathbf{e}_1), (\lambda_2\mathbf{e}_2), \ldots, (\lambda_p\mathbf{e}_p)$ pairs of eigenvalues, eigenvectors derived from a $p \times p$ covariance matrix, $\boldsymbol{\Sigma}$

λ_v maximum eigenvalue

$\mathbf{e}_v$ eigenvector representing the major axis

Y the set of projected data points, $Y = \{y_1, \ldots, y_n\}$

y_q a projected one-dimensional data point, $y_q \in Y$

C the set of projected statistical distance values, $C = \{c_1, \ldots, c_n\}$

c_q the statistical distance of y_q and where $c_q \in C$

H_L the set of m intervals on the left side of a distribution, $H_L = \{h_{L_1}, h_{L_2}, \ldots, h_{L_m}\}$

H_R the set of m intervals on the right side of a distribution, $H_R = \{h_{R_1}, h_{R_2}, \ldots, h_{R_m}\}$

h_{L_i} an interval from the set, H_L

h_{R_i} an interval from the set H_R

$C^{(i)}$ a set of elements (i.e statistical distances) in the interval h_{L_i}, where $c_b^{(i)}$ is the b^{th} element of the total d elements, $\{c_1^{(i)}, c_2^{(i)}, \ldots, c_d^{(i)}\}$ and where $i-1 \leq c_b^{(i)} < i$

$C^{(j)}$ a set of elements (i.e statistical distances) in the interval h_{R_i}, where $c_b^{(j)}$ is the b^{th} element of the total d elements, $\{c_1^{(j)}, c_2^{(j)}, \ldots, c_d^{(j)}\}$ and where $j-1 \leq c_b^{(j)} < j$

$Y^{(i)}$ a set of data points associated with $C^{(i)}$, where $y_b^{(i)} < 0$

$Y^{(j)}$ a set of data points associated with $C^{(j)}$, where $y_b^{(j)} > 0$

$max(C^{(i)})$ the element in $C^{(i)}$ that is associated to the point furthest from 0

$\ulcorner max(C^{(i)}) \urcorner$ the ceiling function of $max(C^{(i)})$

h_{L_m} left furthest interval from the centre point 0 and where $m = \ulcorner max(C^{(i)}) \urcorner$

h_{R_m} right furthest interval from the centre point 0 and where $m = \ulcorner max(C^{(j)}) \urcorner$

$|H_F|$ number of intervals, where $F = L \vee R$, (L=left, R=right)

$|C^{(\tau)}|$ total number of points in the intervals

$\sqrt{\lambda_v}$ standard deviation of the major axis

Ψ cluster density

R total number of ordered cluster (that has been ordered in ascending order)

r index value of an ordered cluster

Ψ^r an ordered cluster

$\{\Psi^r, \ldots, \Psi^R\}$ a set of ordered clusters

G the set of potential threshold values, $G = \{g_{r+1}, \ldots, g_{R-1}\}$

g_i an element from the set G

o a cluster that has not been eliminated and has potential outliers

O the total number of potential outlier clusters

$\neg o$ a cluster with no outliers and has been eliminated

$\neg O$ the total number of clusters containing no outliers and have been eliminated

Δ distance measure

$\neg\eta$ the nearest cluster that has no outliers which has been eliminated

$\bar{\mathbf{x}}_o$ vector of mean values of a cluster which has not been eliminated

$\mathbf{X}_{\neg\eta}$ the set of points from the nearest eliminated cluster that has no outliers, $\mathbf{X}_{\neg\eta} = \{\mathbf{x}_{\neg\eta_1}, \ldots, \mathbf{x}_{\neg\eta_n}\}$

PENGESANAN TITIK TERPENCIL BOLEH SUAI UNTUK TITIK SERAKAN MODEL TAK TENTU

ABSTRAK

Pengesanan titik terpencil adalah proses pengenalpastian corak luar biasa dalam data. Kajian ini memperkenalkan satu kaedah baru untuk mengesan titik terpencil yang terdapat dalam data serakan multivariat yang mana titik terpencil adalah terdiri daripada titik-titik yang berada jauh daripada majoriti titik. Antara cabaran dalam pengesanan titik terpencil adalah kesukaran untuk menentukan taburan bagi memodelkan suatu data serakan. Ini disebabkan oleh ciri-ciri tertentu yang sememangnya telah wujud dalam data itu sendiri, misalnya kepencongan dan kurtosis. Disebabkan ciri-ciri ini, adalah agak mustahil untuk menentukan dengan betul model taburan tanpa sebarang pengetahuan sedia ada ataupun input pengguna. Keadaan ini bertambah teruk apabila data adalah multivariat yang mana akan menyebabkan titik serakan tidak dapat diteliti secara visual. Satu lagi kekangan yang lazim dilihat dalam teknik-teknik sedia ada ialah keperluan untuk mendapat input-input tepat bagi pelbagai parameter. Contoh-contoh parameter ini adalah seperti parameter-parameter fungsi kernel, bilangan kelompok yang ingin dikenal pasti dan nilai-nilai ambang. Kesemua parameter ini sebenarnya boleh mempengaruhi hasil akhir pengesanan ini. Justeru, jika input-input yang diberi adalah salah, hasil akhir yang diperoleh boleh menjadi tidak tepat. Objektif utama kajian ini adalah untuk mencadangkan suatu kaedah tanpa pengawasan yang boleh mengesan titik terpencil dalam data serakan yang mana coraknya adalah sedemikian rupa sehingga model taburan tidak dapat ditentukan dengan mudah. Objektif yang kedua pula bertujuan untuk mewujudkan suatu kaedah boleh suai dengan bilangan parameternya dikurangkan dan input yang mudah ditentukan. Dengan tercapainya objektif-objektif ini, titik terpencil dapat dikesan secara pintar. Dalam kaedah yang dicadangkan, data akan dikelompokkan bagi tujuan mengesan dan menghapuskan kelompok yang tumpat yang biasanya tidak mengandungi titik terpencil. Bagi mengenal pasti kelompok tumpat, ketumpatan setiap kelompok akan ditentukan menerusi suatu kaedah pengiraan yang baru. Kelompok tumpat ini kemudiannya akan dibeza-bezakan dari kelompok yang kurang tumpat menerusi suatu teknik boleh suai, tanpa memerlukan input pengguna. Kesemua langkah ini akan dilakukan secara berulang, sehingga apa yang masih kekal adalah kelompok-kelompok yang kurang tumpat, yang mana titik terpencil berpotensi terkandung di dalamnya. Kemudian, untuk mengesan titik terpencil yang sebenar, beberapa pengiraan kedekatan titik akan dijalankan. Bagi mengesahkan ketepatan dan kecekapan pengesanan, beberapa penilaian telah dijalankan dengan menggunakan beberapa data serakan yang corak taburannya adalah berbeza. Keputusan kajian berkenaan ketepatan yang diperoleh, adalah memberangsangkan. Berbanding dengan teknik-teknik sedia ada, skor F_1 yang diperoleh oleh kaedah yang dicadangkan didapati lebih tinggi sekurang-kurangnya sebanyak 55.6%. Kaedah yang dicadangkan telah juga berjaya digunakan dalam pengesanan anomali imej.

AN ADAPTIVE OUTLIER DETECTION FOR SCATTER POINTS OF UNASCERTAINED MODELS

ABSTRACT

Outlier detection is the identification of unusual patterns in data. This research presents a new method of detecting outliers found in multivariate scatter data, where outliers are those points that lie far away from the majority of points. One of the challenges in outlier detection is the difficulty of determining the distribution to model a scatter data. This is due to the data's certain inherent characteristics, for example, its skewness and kurtosis. Owing to these characteristics, it is therefore quite impossible for the right distribution model to be determined without any prior knowledge or user input. This problem aggravates when data are multivariate, where the scatter of data points cannot be visually inspected. Another problem commonly seen in existing techniques is the need of having precise inputs for various parameters. Examples of these parameters are the parameters of a kernel function, the number of clusters to be identified and threshold values. These parameters are known to have much influence on the detection's final outcome. Thus, if incorrect inputs were given, the final outcome can be very inaccurate. The main objective of this research is to propose an unsupervised method that detects outliers in scatter data where the patterns are such that the distribution model is not easily ascertained. The second objective is to have a method that is adaptive, with the number of parameters reduced and with easy to determine input values. Through having these objectives met, an intelligent way of detection can be better achieved. In the method proposed, data are clustered for the purpose of detecting and eliminating dense clusters which usually do not contain outliers. To identify dense clusters, the density of each cluster is determined through a new method of computation. The dense clusters are then differentiated from the sparse ones through an adaptive technique, with no user input required. All these steps are performed iteratively, till what remains are the sparse clusters which may potentially have outliers. Then, for the true outliers to be finally detected, several point proximity computations are carried out. To verify the detection's accuracy and efficiency, several evaluations were performed using several scatter data that are differently distributed. The results obtained pertaining to accuracy was especially favourable. Compared to existing methods, the F_1 score of the proposed method has shown to be higher at least by 55.6%. The proposed method has also been successfully applied in detecting image anomalies.

CHAPTER 1

INTRODUCTION

1.1 Outlier detection in general

Outlier detection describes the process of identifying unusual patterns in data that do not conform to the normal expected behaviour. This is an area that has been widely researched and its importance is evident through its wide usage seen in various application domains. In different domains, outliers could be termed differently. For example, it could also be known as anomalies, peculiarities, contaminants, errors and etc. Some of the well-known application domains which outlier detection has been applied to are fraud detection in banking and finance, image anomaly detection in industrial quality checks, anomalous patterns in patient medical records, machinery fault detection and abnormalities in surveillance activities. Outliers could be caused by various factors and these include malicious activities, instrumentation errors, natural causes, environmental changes and human errors.

In this research, the outliers specifically refer to data points that exist in scatter plots of multivariate data sets. The detection of such outliers, in simple terms, means the detection of data points that lie far away from the majority of other data points. These data points which are outliers would usually have an inconsistent behaviour or pattern compared to the non-outliers.

The outlier detection techniques or methods that have been developed so far are based on supervised, semisupervised and unsupervised learning approaches. The main difference of the unsupervised approach from the other two approaches is that it does not require a training data set that has labelled data points. The various existing outlier detection methods have also been grouped based on several categories and they are, for example, clustering based techniques, statistical based techniques, spectral based techniques and etc. These categories and the related techniques would be discussed further in the next chapter.

This research proposes an unsupervised outlier detection method where the outliers are detected in a global context. This means the outlyingness of a point would be measured with respect to the whole population of data points in a scatter. The method proposed also has less dependency on user input, thus making it an intelligent and adaptive technique. In the following sections, the problems found in existing approaches are defined and explained. This is followed by the research objectives, the research motivation and a brief description on the proposed method. The final section then describes the contents of each chapter of this thesis.

1.2 Problem definition

This research is based on two areas of the problem that are found in existing methods. The following describes both areas of the problem:

(i) One of the main challenges in outlier detection is the need for prior knowledge of the appropriate distribution to model the data set. However, this often is not possible due to the lack of understanding concerning the data's scatter pattern. When scatter pattern is unknown or remains ambiguous, detection methods based on parametric techniques are no longer suitable for use. This is because the requirement of having to specify the underlying distribution a priori can no longer be met. As for nonparametric techniques, even though techniques such as kernel densities estimation are able to model scatter having unknown density, it however still requires prior knowledge before the type of kernel function can be determined. Although, there are existing techniques where the selection of kernel function has been automated (Howley and Madden, 2006; Ali and Smith-Miles, 2006), they nevertheless do not function in an unsupervised manner. As for the non-statistical techniques, even though such prior knowledge is not required, they are however known to be rather reliant on various parameters in order to perform optimally.

(ii) Most techniques, whether statistical or non-statistical, require exact inputs for various parameters. For such parameters, having precise inputs are highly necessary due to the high influence that they have on the outcome of the outlier detection. In other words, the values given as inputs are critical to the detection's performance. However, determining the values is rather a complex task. For example, determining the value of the kernel function's smoothing parameter, known as the bandwidth, requires additional techniques in the form of data dependent rules to be specified. As for parameters that belong to non-statistical techniques, for example, the number of clusters, the number of nearest neighbours and threshold parameters, it is known that it only takes a slight change in these values, for the outcome generated to end up very differently. In other words, it is rather challenging for these values to be correctly defined.

1.3 Research objectives

This research generally attempts to overcome the problem and limitations stated in Section 1.2. The following are the objectives of this research:

(i) To create an unsupervised method that detects outliers in multivariate scatter data, having patterns such that the appropriate distribution model cannot be easily ascertained.

(ii) To develop a method that is adaptive, with the number of parameters reduced and input values that are easy to determine.

1.4 Research motivation

The objectives mentioned in Section 1.3, are motivated by certain *specific* problems. These problems which this research attempts to solve are explained below:

(i) In this research, the inability to ascertain a suitable model is due to the unavailability of prior knowledge or user input concerning the inherent characteristics of the scatter data's distribution. The characteristics here refer to the different levels of skewness and kurtosis that each different dimension has. The above-mentioned inability to ascertain a model actually further aggravates when data are multivariate since the pattern of scatter can no longer be determined through visual inspection. This research basically attempts to enable outliers to be detected in an unsupervised intelligent manner in spite of the unknown characteristics of the scatter patterns.

(ii) In this research, important parameters are basically parameters that require **exact and specific** values of which must be specified accurately as otherwise, the detection results would be affected. As a solution to this problem, this research attempts to reduce the number of important parameters involved. Secondly, it will also simplify the manner of having the input values determined. In other words, the initial values for these parameters would only require approximate values to be defined. It is sufficient that only approximates be used as these parameters have the ability to adapt according to the current state of the process that is being carried out.

1.5 Brief description of the proposed method

The above-mentioned objectives are achieved through a new way of computation that gauges cluster density, through an intelligent technique in setting a threshold value and also through the enhancements of several existing techniques. In this method, data points are grouped into clusters for the purpose of detecting and eliminating dense clusters based on the notion that outliers are unlikely to be found in denser areas of the scatter.

The clustering is based on an existing technique, a highly efficient Dirichlet process based algorithm. The density of each cluster is then measured through the new clustering density computation. As for the task of identifying the cluster to be eliminated, the proposed method uses a threshold where the value is set through an adaptive technique that does not require any user input. Both the clustering and elimination are done iteratively. Once the data set is left with only a few sparse clusters which potentially contain outliers, the true outliers are then identified through several point proximity computations. Figure 1.1 is a flowchart illustrating the overall idea of the proposed method.

The overall idea underlying the contribution, to the best of our knowledge, is new and has never been developed. Note that, the scope of this research is specially focused on a method that solely detects outliers as the final outcome without considering the detection of clusters. Besides that, it also focuses on a method that detects relatively extreme outliers in a multivariate data set. The multivariate data set as mentioned earlier are such that each dimension has a different level of skewness and kurtosis. All of these would be discussed in detail in the later chapters.

1.6 Organization of thesis

In Chapter 2, the various categories of outlier detection are described. Examples of several existing techniques along with their advantages and limitations are described in detail. At the end of this chapter, a reemphasis on the scope of the proposed method with further details are given.

Chapter 3 gives a detailed description of the contribution. The chapter begins with an overview of the proposed method. Then, in each section, which is dedicated to a particular contributed technique, a detailed explanation is given on the steps involved and also the algorithm.

Chapter 4 consists of the detailed analyses pertaining to the new computational method explained in Chapter 3. These analyses would give one a better understanding

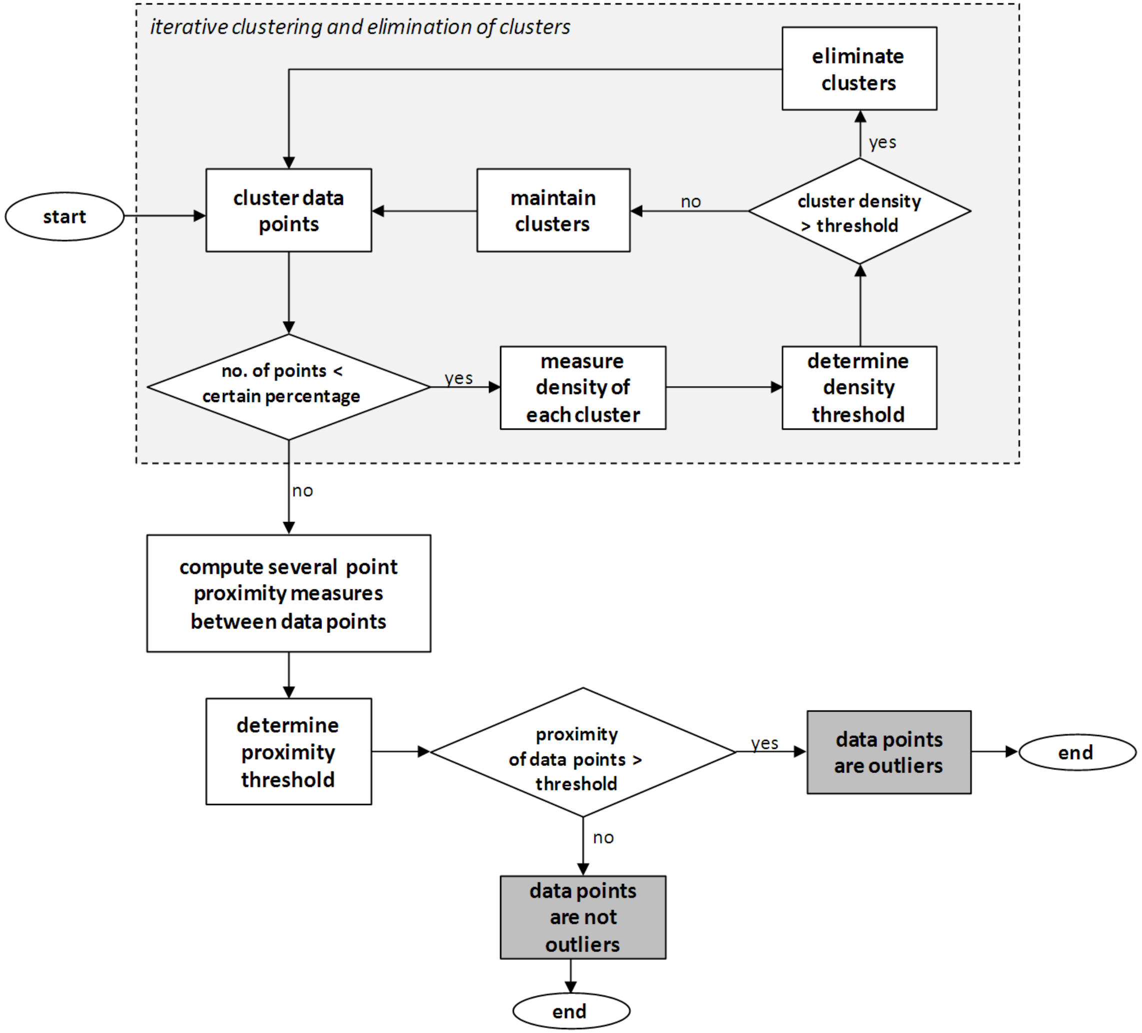
iterative clustering and elimination of clusters
start
cluster data points
maintain clusters
no
cluster density > threshold
yes
eliminate clusters
no. of points < certain percentage
yes
measure density of each cluster
determine density threshold
no
compute several point proximity measures between data points
determine proximity threshold
proximity of data points > threshold
yes
data points are outliers
end
no
data points are not outliers
end

Figure 1.1: A flowchart illustrating the overall idea of the proposed method.

on the necessity of having certain parameters and certain steps as part of the proposed computation.

Chapter 5 provides a detailed explanation on the different types of evaluations that have been carried out along with the related results and discussions. The first evaluation is related to the several important functionality of the proposed method. This is followed by the evaluation concerning the outlier detection accuracy and its computational cost. These results obtained which have been compared against the results of other methods are also described. Besides that, the usefulness of the proposed method is also demonstrated through its application in anomaly detection in textured images.

The final chapter, Chapter 6, concludes the thesis by discussing general issues of this research and future works. This chapter also revisits the various contributions with a brief explanation given for each contribution.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter describes the different types of outlier detection methods. The various existing works in the literature can be divided into several categories. In this chapter, these categories are described along with examples of the existing methods, their advantages and limitations. At the end of this chapter, the considerations that were taken in designing the proposed method are discussed.

2.2 A brief introduction to outlier detection

The research on outlier detection began as early as the 19th century where it was a subject mainly studied by the statistics community. Over the years, various detection techniques were developed and improved by different research communities. The outlier detection techniques developed were mostly able to be translated into a contribution significant to an application domain. Although an outlier detection technique is most often applied to a specific application domain, there are also generic ones that may be applied to several different application domains.

Outliers should not be confused with noise and it is not related to the study of noise removal (Teng et al., 1990) that deals with unwanted noisy data. Noise may be generally defined as an occurrence that hinders data analysis and therefore requires removal. Outliers on the other hand are regarded as occurrences of interest and of relevance to be analysed.

2.3 Factors influencing the formulation and algorithm of an outlier detection technique

The specific formulation or the algorithm of an outlier detection technique is based on factors such as *type of outliers* and *labels* which are described below:

(a) **Type of outliers**

Outliers could be further defined as either point outliers, contextual outliers or collective outliers. Point outliers may be determined by measuring its extremeness with respect to the rest of the population data (Chandola et al., 2009). As for contextual outliers, the consideration of whether or not a data point is an outlier is based on a given specific context. To illustrate the meaning of specific context, let's say if the temperature reading of 10°C, taken during winter at a certain place is considered an occurrence that is usual. Then, this same value is nevertheless an outlier, if it had occurred during the summer. Thus, the specific context here means the specific season. The last type of outliers, known as collective outliers refers to the case where an individual data point would only be considered an outlier if it is found to be in a group of outlying data points. Otherwise, if it is by itself, it is not an outlier.

(b) **Labels**

Labels are indicators that denote whether or not a data point is an outlier. Obtaining accurately labelled data is known to be time consuming and labour intensive as most often, labelling of data has to be done manually by a human expert. There are three modes of data labelling and they are: supervised, semisupervised and unsupervised. The various outlier detection techniques in the literature are associated with either one or more of these three modes. The techniques associated with supervised mode are usually dependent on training data set that has labelled data points (Tan et al., 2005). Through the labels obtained from the training data, outliers are then detected from the test data sets of which the existence of outliers is initially unknown. As for the semisupervised mode, the techniques related are similar to those that belong to the supervised mode except that the labelling of data is done only for non-outlier data points of the training data set. Then, finally, the outlier detection techniques that are associated with the unsupervised mode are basically techniques where labelling of data is not required at all.

2.3.1 The type of outlier detected by the proposed method

Of the different types of outliers discussed above, the type of outlier the proposed method is specially focused on is **point outliers**. Only one type is considered because attempting to detect all types while ensuring the objectives in Chapter 1 are met would result in a research that is too complex and a scope too large.

As described in Chapter 1, the method proposed is an unsupervised detection that does not require the labelling of data. Nevertheless, it should once again be noted that the proposed method also operates intelligently and is adaptive. This is because more than just being independent from needing training data sets or labelled data, the method proposed has the capability to work without requiring prior knowledge concerning the inherent characteristics of scatter data's distribution and also with a reduced number of parameters which only require approximate values. However, due to the complexities that naturally occur in methods with such capabilities, is the reason that the method proposed will only be used for the detection of point outliers. The positions of these point outliers on the scatter plot are rather obvious and this is due to them having positions that are relatively far from the rest of the scatter population.

Now, since among all the points on the scatter plot, the distance of an outlier point from the other points has been considered as being relatively much further than the distance among non-outlier points, the outlier has therefore been given a specific term known as **relatively extreme outliers**. However, relatively extreme outliers should not be confused with another term that is known as *extreme outliers*. The latter, is a well known term that defines the position of an outlier as being three times the size of the interquartile range, measured from either the upper or lower quartile and where these quartiles are measured from an ordered set of numbers (Devore and Berk, 2007). Note that, the relatively extreme outliers that are detected by this research's proposed method do not follow the definition that has been given for extreme outliers.

Lastly, it should be known that although the occurrence of these relatively extreme outliers on the scatter may be obvious to humans, it is nonetheless still a challenge for the computer to detect them without being too reliant on user input. Detecting such outliers in a manner that is intelligent and adaptive is what this research strives to achieve.

2.4 The categorization of outlier detection techniques

In the following subsections, the various categories of outlier detection are discussed.

2.4.1 Nearest neighbour based techniques

Nearest neighbour based outlier detection techniques assume that non-outliers should be found in dense neighbourhoods, while outliers are located far from their closest neighbour. Usually, to detect outliers, distance or similarity measure is made between two data points. One of the more popular ways to measure distance or similarity is the Euclidean distance (Tan et al., 2005, Chap. 2). Techniques of this category can be further divided into two sub-categories and they are:

(i) Techniques where the outlier detection is based on distance that is measured between a data point and the k^{th} nearest neighbour.

(ii) Techniques where the outlier detection is based on each data point's relative density. A data point is an outlier if it lies in a neighbourhood which has low density. If a data point lies in a dense neighbourhood, it is not an outlier.

Described as follows are examples of the related works. The related works of the first sub-category are described first followed by those from the second sub-category.

Eskin et al. (2002), Angiulli and Pizzuti (2002) and Zhang and Wang (2006) are examples of related works from the first sub-category. They can be generally described as techniques that compute the outlier score based on sum of distances of a data point from its k^{th} neighbour. Besides this approach, the outlier score of data point may also be derived by counting the number of nearest neighbours that are not more than a certain distance compared to the given data item (Knorr and Ng, 1997, 1998), (Knorr et al., 2000). There are also contributions that are focused on improving the efficiency of the nearest neighbour technique. One example is a technique by Wu and Jermaine (2006) which had proposed a sampling based technique. Instead of involving every data point in the data set, they proposed that the nearest neighbour computation should be based on a smaller set of data samples. Apart from these techniques, there is also one that is based on the partitioning and pruning approach (Ramaswamy et al., 2000; Ghoting et al., 2008; Tao et al., 2006). For example, Ramaswamy et al. (2000) proposed a technique where first, data partitioning or clustering is performed. Then, the k^{th} nearest neighbour for each point in the cluster is measured. From the distance measured and also the minimum bounding rectangular condition, the upper and lower bounds are calculated. The values obtained would determine the outlier threshold value. Then for those partitions that have upper bounds below the threshold, they will be pruned, as most likely they do not contain outliers. Finally, through an index based algorithm, a pre-specified number of outliers are detected. Basically, outliers are those points with the highest distance value.

A very well known technique of the second sub-category is the Local Outlier Factor (LOF) (Breunig et al., 2000). The LOF score for any given data point is equivalent to the ratio of the local density of a data point and the average local density of the data point's k^{th} nearest neighbours. In order to identify the local density of a data point, a hyper-sphere containing the data point's nearest neighbours is identified. The local density is calculated as the reciprocal of the average distance to k nearest neighbour. For a data point to be outlying, one can expect that its local density will be lower than that of its nearest neighbours whereas an outlier will have a similar density as its nearest neighbours. There are several variations to the LOF and where one of them is known as the Connectivity-based Outlier Factor (COF) (Tang et al., 2002). The difference between the COF and LOF is basically the manner in which the k^{th} neighbourhood of a data point is computed. For COF, a data point would only be added to a neighbourhood if the distance measured between them obtains a minimum value. Data points are then added this way until the neighbourhood has reached the pre-specified k number of data points. Nevertheless, the manner in which the outlier score is computed for COF is the same as that for LOF. Other examples of techniques that are also variants to the LOF is the approach that detects spatial anomalies or outliers in climate data (Sun and Chawla, 2004; Chawla and Sun, 2006) and also the approach that had used similarity measure to handle data with categorical attributes (Yu et al., 2006).

The main advantage of nearest neighbour based techniques:

These techniques do not require any assumption to be made concerning the distribution.

The main disadvantages of nearest neighbour based techniques:

(i) The measure of distance has to be calculated for every single data point. Thus, computation complexity could be rather significant.

(ii) The techniques are biased to the value k. Different k can give very different results.

2.4.2 Clustering based techniques

Clustering may be described as a process that divides data into groups of clusters. Clustering is generally an unsupervised technique although lately, semisupervised clustering (Basu et al., 2004) has been explored as well. There are several assumptions related to clustering based outlier detection techniques. In the ensuing paragraphs, the assumptions and the related works are described.

The first assumption states that outliers are data points that do not belong to any cluster. An example of clustering algorithms that are based on this assumption is FindOut algorithm (Yu et al., 2002) which is an extension of the WaveCluster algorithm (Sheikholeslami et al., 1998) where the detected clusters are removed from the data and the remaining data points are declared as outliers.

The second assumption states that non-outliers are data points that are positioned near to the closest cluster mean or cluster centroid whereas outliers are those positioned further away. Basically, clustering algorithm of this assumption involves two steps. The first step involves clustering of data through a clustering algorithm. The second step calculates the outlier score based on distance measured for each outlier data point to the closest cluster centroid.

K-means algorithm (Hartigan and Wong, 1979; Telgarsky and Vattani, 2010), and the Expectation-Maximization (EM) algorithm (Hartley, 1958; Dempster et al., 1977), are two examples of clustering algorithms used in outlier detection and where the detection is accomplished through the second step stated above. Clustering performed through *K*-means would result in the partitioning of data points into *K* clusters where the assignment of a data point to a cluster is based on its distance to the nearest cluster mean. Koupaie et al. (2013), is one example of an outlier detection method that had used K-means clustering, with the inclusion of some adaptations. Then, as for the EM algorithm, the assignment of a data point is based on the probability a data point belongs to a distribution. The computation of this probability, known as the Expectation step (E step) is performed for every data point. Then, the Maximization step (M step) performed thereafter, computes the new parameter estimates for each distribution. Both the E and M steps are performed iteratively until the distribution parameters reach convergence. Goldstein (2012) is one example of an EM algorithm based outlier detection method, that has some adaptations included. Note that, both the K-means and EM algorithm along with most other clustering algorithms, require the number of clusters to be pre-specified.

As for the third assumption related to clustering, the assumption states that non outliers are data points found in large dense clusters whereas outliers are found in small sparse clusters. He et al. (2003) had proposed a technique known as FindCBLOF that assigns an outlier score known as Cluster-Based Local Outlier Factor (CBLOF) for each data point. From the score obtained, one will know the size of the cluster which the data point belongs to and the distance of the data point to the cluster centroid. Another example is a technique by Wang and Dunson (2011), which proposed a clustering algorithm for mixture models based on Dirichlet process (Antoniak, 1974), which is known for not requiring user input or the use of model selection techniques in determining parameter values such as the number of clusters (Orbanz and Teh, 2010). Hence, the outcome of the clustering will be based on the data's scatter pattern and not the parameter that defines the number of clusters, of which may have the possibility of being given an incorrect value (Teh et al., 2006). Having an incorrect value as input to this parameter would certainly affect the accuracy of clustering outcome as this parameter is crucial in determining the final clustering output. The algorithm that was proposed by Wang and Dunson (2011) is called the Sequential Updating Greedy Search (SUGS) algorithm. The SUGS algorithm creates a mixture model on the clusters that were created based on the product of the prior probability representing the clusters of data points and the prior probability of parameters of each cluster. These clusters are formed through the adding of data points that is done one at a time in a sequential manner. The cluster to which a data point was added to, is the cluster which had maximized the posterior probability. The posterior probability here is specifically the conditional posterior probability of a cluster given its data points. From the experimental results, the SUGS has shown itself to be successful in modelling one sided long-tailed, one-dimensional distribution. It is indeed common for such distribution to have outliers. The SUGS is also proven to be a much faster technique compared to other Dirichlet process based techniques such as the Markov Chain Monte Carlo (MCMC) simulation based methods (Neal, 2000; West and Escobar, 1994) and Variational Bayes (Blei and Jordan, 2006). MCMC based methods are known to be inefficient when data sets are large or high dimensional.

The main advantage of clustering based techniques:

Techniques of this category can operate under unsupervised mode where labels are not required.

The main disadvantages of clustering based techniques:

Almost all clustering techniques except for those that are Dirichlet process based, require the number of clusters to be specified. In order to have a correct input which would give a correct outcome, a good understanding of the application domain is usually necessary besides making use of model selection techniques. Without these, the results could turn out highly inaccurate.

As for techniques that are based on the second assumption, they usually require distance to be measured for every data point to its closest centroid. This step could be computationally intensive when data sets are large or high dimensional.

2.4.3 Statistical based techniques

For statistical based outlier detection techniques, outliers are determined based on how likely it belongs to a distribution. Techniques of this category are based on an assumption that non-outliers are data points that occur in high probability regions of a distribution model while outliers occur in low probability regions. Statistical based techniques would basically attempt to fit a distribution model to a given data set. Both parametric and nonparametric techniques have been applied to fit a statistical model. Below is the discussion of both these techniques.

2.4.3(a) Parametric techniques

Most parametric techniques are based on Gaussian models where it assumes that data are generated from a Gaussian distribution. The simplest way of detection is to declare all data points which are more than three standard deviations 3σ from the mean μ as outliers. Yet another approach is through the calculation of score based on the distance of a data point from the mean of the distribution. A threshold is then applied to the score to determine the outliers. The more sophisticated technique and which is rather well known however is a technique by Rousseeuw and Driessen (1999). The technique which they contributed is an algorithm that detects outliers through a minimized covariance determinant and thus, this technique is known as MCD (Minimum Covariance Determinant). The method proposed is a highly robust estimator that can be applied to multivariate scatter data. The objective of MCD is to find a certain number of observations, *i* from the total *n*, that will finally have a covariance matrix with the lowest determinant. The MCD estimate of location may be determined by computing the average of these *i* data points. Performing these steps, however, has actually been found to be computationally difficult. Thus, a fast algorithm known as FAST-MCD was proposed. This algorithm is based on a procedure of generating initial estimates known as C-Step. Other than that, the algorithm also consists of two techniques known as selective iteration and nested technique. Through selective iteration, the number of C-step is reduced and thus, the speed of the algorithm is also improved. As for the nested technique, C-step will be carried out on several nested random subsets of data points instead of the entire data set.

2.4.3(b) Non-parametric techniques

Outlier detection techniques which are based on nonparametric statistical models, do not require statistical or distribution models to be defined a priori. Instead, the models are determined from the given data. Earlier works in this area were mostly based on histograms which were applied to applications such as network intrusion detection (Ho et al., 1999; Yamanishi and Takeuchi, 2001; Yamanishi et al., 2004) and Web-based attacks detection (Kruegel et al., 2002; Kruegel and Vigna, 2003).

An example of a simple but rather effective outlier detection for univariate data is the method by Laurikkala et al. (2000) that uses box-plots. It was mentioned that multivariate outliers are not necessarily univariate outliers. Thus, the box-plots which have been plotted for each dimension, may not be effective in detecting outliers. To detect multivariate outliers, it was proposed that Mahalanobis distance be measured for every data point before representing them in a box-plot. However, it was stated that using Mahalanobis distance does have a limitation where it is more suited when the data is normally distributed.

The more sophisticated non parametric techniques however would be those that are based on kernel function. Outlier detection techniques based on kernel functions are quite similar to the parametric techniques described earlier. The difference however is the density estimation technique used. An example of a non-parametric technique that is based on the kernel function is the Parzen windows estimation (Parzen, 1962). A common application domain which this technique has been applied to is the detection of network intrusion (Yeung and Chow, 2002).

One of the more recent non parametric techniques in the literature is a technique that is based on Kernelized Spatial Depth (KSD) (Chen et al., 2009). With the right choice of positive definite kernel, KSD is able to capture the pattern of the data scatter. Outliers are then identified through the KSD-based algorithm which computes the spatial depth value. If the value obtained for a data point is less than a spatial threshold, the data point would be considered as an outlier.

Another kernel based method for outlier detection is a method proposed by Latecki et al. (2007) where outliers are detected by comparing the local density of every data points to the average density of its neighbours. The density of each of its neighbours is estimated through a nonparametric kernel estimate. The Gaussian kernel function which is the author's choice of kernel function had been enhanced where the values of the exponent are based on the Mahalanobis distance between data points and its neighbourhood.

The main advantage of statistical based techniques:

Techniques of this category are accurate if the distribution of the model is known.

The main disadvantages of statistical based techniques:

Parametric techniques are usually heavily dependent on the assumption that data are generated from a particular distribution model. Most often this cannot be ascertained and is especially true if data are high dimensional. As for nonparameteric methods, the choice of kernel function and the bandwidth parameter values are important as both have a high influence on the results of the density estimate. Making the right choice, however, can be difficult when no prior knowledge regarding the data set's underlying distribution is available.

2.4.4 Spectral based techniques

Generally, spectral based outlier detection techniques are based on the assumption that data can be projected into a lower dimensional space in which both non-outliers and outliers will appear significantly different. This is because non-outlier data points are usually correlated while outliers would deviate from the correlation structure. The technique proposed by Dutta et al. (2007) had adopted an approach based on Principal Component Analysis (PCA) (Jolliffe, 2002) to detect anomaly or outliers in astronomy catalogues. Ide and Kashima (2004), on the other hand, proposed a technique that detects outliers in time series of graphs. Sun et al. (2008) had also contributed a technique for time series of graphs. The authors had proposed a Compact Matrix Decomposition (CMD) that is performed on a sequence of graphs in which the outlier will be detected. Then, Shyu et al. (2003) had proposed an outlier detection technique where robust PCA is performed. Firstly, the principal components from the covariance matrix of the non-outlier training data is estimated. The next step involves the comparison of each data point with components. An outlier score is then assigned based on the distance between the data point and principal component.

The main advantage of spectral based techniques:

Since techniques of this category basically perform dimensionality reduction, they are therefore suitable for handling data with high dimensionality. Often, spectral based techniques act as a preprocessing step which is followed by any existing outlier detection technique.

The main disadvantage of spectral based techniques:

Spectral-based techniques are only useful if the outliers and non-outliers are separable after being projected to the lower dimension.

2.4.5 Classification based techniques

Classification based outlier detection techniques classify data points that are from test data sets as either outliers or non-outliers based on the labelled data obtained from the training data sets. Since, in this thesis, the method proposed does not depend on labelled data, the techniques of this category can be therefore be viewed as less relevant. Due to this reason, only a brief description is given concerning the related techniques of this category.

Classification based techniques can be grouped into two categories which are multi-class and one class. Techniques based multi-class assume that the training data contains labelled data points that belong to multiple classes. Techniques based on one-class assume that the training data has only one class label.

Support Vector Machines (SVMs) (Vapnik, 2000) have been applied in outlier detection in a one-class setting. There are a few variations of one-class SVM and the one by Schölkopf et al. (2001), trains the hyperplane that has the maximum distance between the data points and the origin of the coordinate system. Then based on a certain percentage, the outliers are detected from the rest of the data (Hejazi and Singh, 2013). Outlier detection based on SVM has also been applied to detect system call intrusions (Eskin et al., 2001; Heller et al., 2003; Lazarevic et al., 2003) and to detect outliers in audio signal data (Davy and Godsill, 2002).

To detect outliers in multi-class settings, techniques based on Naive-Bayesian have often been used. There are several variants that have been applied to different types of application domains. For example, network intrusion detection (Barbara et al., 2001; Bronstein et al., 2001; Sebyala et al., 2002; Valdes and Skinner, 2000), and outlier detection in text data (Baker et al., 1999).