2.1 Introduction

This chapter considers previous research done in the field of outliers specification in data mining techniques.

The field of data mining has been studied extensively, most of the work has concentrated on discovery of patterns Mansur et al. (2005). Outlier detection as a branch of data mining has many important applications, and deserves more attention from data mining community Fayyad et al. (1996). Most methods in the early work that detects outliers independently have been developed in field of Statistics. Finding , removing and detecting outliers is very important in data mining, for example error in large databases can be extremely common, so as an important property of a data mining algorithm is robustness with respect to outliers in the database.
Most sophisticated methods in data mining addressed that this problem to some extent, but not fully, and could be improved by addressing the problem more directly John (1995). The identification of outliers could lead to the discovery of unexpected knowledge in areas such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors, discovering computer intrusion, etc. It was focused on outlier identification and provided a description of why an identified outlier exceptional, based on Distance - Based outlier detection and Density-Based outlier detection (Mansur et al. (2005)).

Xi (2008) discussed and compared different outlier detection approach from data mining perspective, which could be categorized into two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which could be grouped into statistical-based approach, distance - based approach, deviation-based approach, density-based approach. The spatial outlier approach analyzes outlier based on spatial dataset that non-spatial and spatial data were significantly different from transaction data, which could be grouped into space-based approach and graph-based approach.

Pasha & Umesh (2013) had been proposed several techniques to target a particular application domain. The classification of outlier detection techniques based on the applied knowledge discipline provided an idea of the research done by different communities and also highlights the unexplored research avenues for the outlier detection problem. Discussed of the behavior of different techniques would be done, with respect to the nature. The feasibility of a technique in a particular problem setting also depended on other constraints.
2.2 Data Mining and Outliers

Many KDD applications, such as criminal activities in e-business environment, finding the rare events or the outliers could be more interesting than finding the common patterns. Recently, many studies had been focused on outlier detection for large datasets (e.g. Breunig et al. (2000), Han et al. (2011), Khan et al. (2014). Basic approaches currently used for solving this problem were considered, and their advantages and disadvantages were discussed by Petrovskiy (2003). In order to solving the many applications of industrial problem for data mining and data features, a clustering based data mining algorithms to solve the issue of tax discovery of outlier data. Liu et al. (2012). Several medical applicaions based support vector machine algorithm was used for experimental analysis Aljumah et al. (2013).

Chandore & Chatur (2013) had dealt the problem with detecting outlier over Dynamic data stream and specific techniques used for detecting outlier over streaming data in data mining. Sreevidya (2014) had discussed the problem with detecting outlier over data stream and specific techniques used for detecting outlier over streaming data in data mining. Various of measures central tendency methods like mean, mode, median and Inter-Quartile – Range(IQR) on real time databases and the experimental results were generated by Sunitha et al. (2014). Singh & Pathak (2014) described an approach which uses Univariate outlier detection as a pre-processing step to detect the outlier and then applied K-means algorithm hence to analyse the effects of the outliers on the cluster analysis of dataset. Lekhi & Mahajan (2015) Proposed method uses hybrid approach to reduce the number of outliers.
2.3 Historical Background of Outliers

Detection of outliers in the analysis of the data sets dates back to 18th century. Bernoulli (1777) pointed out the practice of deleting the outliers about 200 years ago. Deletion of outliers was not a proper solution to handle the outliers but this remained a common practice in past. To address the problem of outliers in the data, the first statistical technique was developed in 1850 (Beckman & Cook (1983)). Some of the researchers argued that extreme observations should be kept as a part of data as these observations provide very useful information about the data. For example, Bessel and Baeuer (1838) claimed that one should not delete extreme observations just due to their gap from the remaining data (cited in Barnett & Lewis, 1978). The recommendation of Legendre (1805) is not to rub out the extreme observations “adjusted too large to be admissible”. Some of the researchers favoured to clean the data from extreme observations as they distorted the estimates. An astronomer of 19th century, Boscovitch, put aside the recommendations of the Legendre and led them to delete (ad hoc adjustment) perhaps favoring the Pierce (1852), Chauvenet (1863) or Wright (1884). Cousineau & Chartier (2010) said that outliers were always the result of some spurious activity and should be deleted. Deleting or keeping the outliers in the data is as hotly discussed issue today as it was 200 years ago. Bendre & Kale (1985), Davies & Gather (1993), Iglewicz & Hoaglin (1993) and Barnett & Lewis (1984) had conducted a number of studies to handle issues of outliers. Defining outliers by their distance to neighboring examples was a popular approach to finding unusual examples in a dataset known to be distance based outlier detection technique. Saad & Hewahi (2009) introduced Class Outlier Distance Bases (CODB) outlier's detection
procedure and proved that it was better than distance based outlier's detection method. Verma (1997) emphasized for detection of outliers in univariate data instead of accommodating the outliers because it provided better estimate of mean and other statistical parameters in an international geochemical reference material (RM).

2.4 Outliers in Data Sets

Quesenberry & David (1961) discussed the rejection and location of outlying observations that there might be several ways of approaching the problem, which depended to a large extent on the object in view. One might be particularly interested in identifying which were the genuinely exceptional observations, in order to create a new insight into the phenomena under study. In the first case, criterion of what was best might be the effect on the standard error of estimation, in the second case of the risk of wrongly deciding whether an observation was exceptional or not. These procedures discussed in the following chapter started from the basis of risks of misclassification rather than of estimation errors. Grubbs (1969) gave the procedures for determining statistically whether the highest observation, the highest and lowest observations, the two highest observations, the lowest observations, or more of the observation in the sample were statistical outliers.

McMillan (1971), Moran & Mcmillan (1973) had the performances of three procedures for the treatments of outliers in normal samples that was evaluated. They had compared three procedures for the detection of outliers in samples from a Normal population. One of these was a sequential application of a maximum residual test. One step in McMillan's derivation of the probability of detecting two outliers by
using this test was incorrect. A correct derivation was presented in the note and some numerical results in McMillan's paper were appropriately amended. McMillan's qualitative statements on the performance of the test were essentially unaffected.

Tietjen & Moore (1972), Tietjen et al. (1973) were focused on the problems of repeated application and "masking". An appropriate to overcome these problems were two new statistics: $L_k$ which was based on the $k$ largest (observed) values and $E_k$ which was based on the $k$ largest (in absolute value) residuals. Tables of approximate critical values for these statistics were given for 0.01, .025, 0.05, and 0.10 levels of significance and for sample size $n = 3(1)20(5)50$. A procedure of studentizing or standardizing the residuals by dividing them by their estimated standard deviations was proposed for testing for outliers in simple linear regression.

Prescott (1975) considered a test statistic for detecting outliers in linear models involving residuals standardized by their individual standard deviations and it was suggested that its critical values were adequately approximated by upper bounds for the critical values of a similar test statistic involving residuals standardized by a constant standard deviation. Two regression analyses were given to illustrate the procedure and did not require a re-analysis with the suspected outlier omitted or treated as a missing value.

Rosner (1975) was concerned with “many outlier” procedures i.e., procedures that could detect more than one outlier in a sample. Several many outlier procedures were proposed via power comparisons were found to be much superior to one outlier procedures in detecting many outliers. Many outlier procedures based on the extreme
studentized deviate (ESD) was slightly the best. Finally, 5%, 1% and .5% points were
given for the ESD procedure for various sample sizes.

Galpin & Hawkins (1981) proposed accurate bounds a represented for the fractiles
of the maximum normed residual (which was often used to test for a single outlier)
for two-way and three-way layouts. It was shown that the second Bonferroni bound
of the critical value, while not conservative, was an excellent approximation to the
critical value, being much more accurate than the first Bonferroni upper bound. The
third Bonferroni (upper) bound, which, although conservative, was expensive to
calculate, agrees with the second bound to at least four decimal places for all factor
combinations considered.

Rousseeuw & Zomeren (1990) had deals with masking problem affected in a
multivariate data, especially when there were several outliers. The classical identification
method did not always find them, because it was based on the sample mean and
covariance matrix, which were themselves affected by the outliers. To avoid the
masking effect, to compute distances based on very robust estimates of location and
covariance. These robust distances were better suited to expose the outliers.

Paul & Fung (1991) proposed procedures for detecting multiple $y$ outliers in linear
regression. A generalized extreme studentized residual (GESR) procedure, which
controlled type I error rate, was developed with several examples. An approximate
formula to calculate the percentiles was given for large samples and more accurate
percentiles for $n \leq 25$ were tabulated. The performance of this procedure was
compared with others by Monte Carlo techniques and found to be superior. The
procedure, however, failed in detecting outliers that were on high-leverage cases. They suggested a two-phase procedure. The phase 1, a set of suspect observations was identified by GESR and one of the diagnostics applied sequentially and phase 2, a backward testing was conducted by using the GESR procedure to see which of the suspect cases were outliers.

Davies & Gather (1993) discussed about the identifying outliers to assume that the outliers had a different distribution from the remaining observations. They defined the outliers in terms of their position relative to the model for the good observations. The Methods based on the robust statistics and outward testing were shown to have the highest possible breakdown points in a sense derived from Donoho and Huber. But a more detailed analysis showed that the methods based on robust statistics performed better with respect to worst-case behavior. M. R. David & Woodruff (1996) had discussed the difficulties occur in the multivariate outliers problem increased with dimension of data and also described significant difference of improvements in the methods for outlier detection with simulation technique with examples.

Linsinger et al. (1998) had found the outlier test procedure to influence the interlaboratory standard deviations (SD), but not the averages. It was shown that even small number of differences in the numbers of outliers detected could change the SD severely. Comparing the outliers test procedures of Hampel, Grubbs and Graf-Henning, it was found that Hampel test detected the most outliers. Penny & Jolliffe (2001) had proposed a clinical trial based new treatment for identifying multivariate outliers.
Southworth (2008) had discussed identifying outliers in clinical trial data, dealing with unmasking multivariate outliers. Bohrer (2008) had using Dixon’s outlier test had been calculated by using Monte Carlo simulation, one sided two-sided case critical values were determined. Barbato et al. (2011) discussed about the several statistical methods that were currently in use for outlier identification and their performance were compared theoretically for typical statistical distributions of experimental data, considering values derived from the distribution of extreme order statistics as reference terms.

Dovoedo & Chakraborti (2013) dealt with the case of multivariate skewed data, specifically when the data followed the multivariate skew-normal distribution. They also compared the outlier detection capabilities of four robust outlier detection methods through a simulation study. Christophe et al. (2013) suggested that standard deviation method was not suitable for outlier detection, for this reason, that it was considered as the poor method. So they used MAD method for robust estimator for the median absolute deviation about the median. Obikee et al. (2014) had compared the several outlier identification techniques that included modified Z-Scores method, which was being used for the simulation study based on the disease group data set generated normal distribution.

### 2.5 Outlier Detection Methods

In this literature, the most important kinds of outlier detection approaches were the following:

1. Distribution-based approaches were used in standard statistical distribution by Šaltenis (2004). They deployed some standard distribution model and recognized
as outliers those points which deviate from the model. A large number of tests were required in order to decide which distribution model fit the arbitrary data set best. Fitting the data with the standard distributions were quite costly.

2. Clustering-based approaches had, as main objective, to discover clusters, and so they were not developed to detect outliers. Clustering was a technique aimed at grouping similar data instances in groups or clusters *Jain & Dubes (1988)*. Although the main objective of clustering was to discover clusters, it had become an important tool for outlier detection and analysis. Indeed, several clustering-based outlier detection techniques had been developed. Most of these techniques rely on the key assumption that normal data instances belonged to large and dense clusters, while outliers form very small clusters or isolated elements.

3. Depth-based approaches based on computational geometry and computed different layers of k-dimensional convex hulls. Outliers were more likely to be data objects with smaller depths. Depth-based approach was also applied for spatial outlier detection *Cárdenas-Montes (2014)*.

4. Distance-based methods used a distance metric to measure the distances among the data instances. Problems might occur if the parameters of the data were very different from each other in different regions of the data set.

5. Density-based approaches applied a local cluster criterion. Clusters were regarded as regions in the data space in which the objects were dense, and which were separated by regions of low object density (outlier). These regions might have an arbitrary shape and the objects inside a region might be arbitrarily distributed.
2.5.1 Distribution based Outlier Detection

Distribution-based methods rely on assumptions that the data followed a statistical distribution model e.g., Normal, Poisson, Binomial. Hence, a point that deviated significantly from the data model was declared as an outlier.

van der Loo (2010) Two univariate outlier detection methods were introduced. In both methods, the distribution of the bulk of observed data was approximated by regression of the observed values on their estimated QQ plot positions using a model cumulative distribution function. Having obtained a description of the bulk distribution, They gave two methods to determine if extreme observations were designated as outliers. In Method I, they determined the value above which less than a certain number of observations (say 0.5) were expected, given the total number of observations and the fitted distribution. In Method II, they devised a test statistic to determine whether an extreme value can be drawn from the same distribution as the bulk data. Both methods have been implemented in the “extremevalues” R package which has been made available via the CRAN web archive. An outlier detection method based Method I using the lognormal distribution had been implemented for the Structural Business Statistics at Statistics Netherlands.

Distribution-based methods were the earliest parametric methods to face the outlier detection problem. As Parametric methods, they directly calculate the parameters of this distribution based on means and covariance of the original data. Then, they employ statistical tests to determine a point as an outlier depending on whether it deviates significantly from the data model.
2.5.2 Depth-based methods

Outlier detection methods that were based on statistical depths had been studied in statistics and computational geometry. These methods provide a center-outward ordering of observations. Each data point was assigned by a depth [100] and outliers were expected to appear more likely in outer layers with small depth values than in inner layers with large depth values. Depth-based methods were completely data-driven and avoid strong distributional assumption.

Moreover, they provide intuitive visualization of the data set via depth contours for a low-dimensional input space. Most of the various depths, spatial depth was especially appealing because of its computational efficiency and mathematical tractability. Spatial depth has been applied in clustering and classification problems. Because each observation from a dataset contributes equally to the value of depth function, spatial depth takes a global view of the data set. Cárdenas-Montes (2014) proposed a new approach aimed to detect outliers in very large data sets with a limited execution time was presented. This algorithm visualizes the tuples as N-dimensional particles able to create a potential well around them. Later, the potential created by all the particles was used to discriminate the outliers from the objects composing clusters.

2.5.3 Graph-based methods

Graph-based methods make use of a powerful tool data image and map the data into a graph to visualize the single or multi-dimensional data spaces. Outliers were those points that were present in particular positions of the graph. These methods
were suitable to identify outliers in real-valued and categorical data. Shekhar et al. (2001) focus on detecting spatial outliers in graph structured data sets. They define statistical tests, analyze the statistical foundation underlying our approach, design several fast algorithms to detect spatial outliers, and provide a cost model for outlier detection procedures. Moonesinghe & Tan (2008) have introduces a stochastic graph-based algorithm, called OutRank, for detecting outliers in data. They consider two approaches for constructing a graph representation of the data, based on the object similarity and number of shared neighbors between objects. The heart of this approach was the Markov chain model that was built upon this graph, which assigns an outlier score to each object. Using this framework, they show that our algorithm was more robust than the existing outlier detection schemes and can effectively address the inherent problems of such schemes. Empirical studies conducted on both real and synthetic data sets show that significant improvements in detection rate and false alarm rate were achieved using the proposed framework.

### 2.5.4 Clustering Based Methods

Traditional clustering based methods were developed to optimize the process of clustering of data, where outlier detection was only by-product of no interest. The novel clustering-based outlier detection methods can effectively identify outliers as points that do not belong to clusters of a data set or as clusters that were significantly smaller than other clusters. Many authors worked in this topic of k-means clustering based outlier detection in data mining. There were some related literatures as follows. Hartigan & Wong (1979) proposed an efficient algorithm of k-means clustering. The
K-means algorithm was to divide M points in N dimensions into K clusters so that the within-cluster sum of squares was minimized. It was not practical to require that the solution has minimal sum of squares against all partitions, except when M, N were small and K = 2. They find "local" optima, solutions such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares.

Gray & Ling (1984) describes a new methodology for the detection of influential subsets in regression. The method was based on an adaptation of computational and graphical techniques used in cluster analysis and makes use of some general properties of influential subsets, but it was independent of any specific measure of influence. For small to moderate data sets the proposed method was computationally efficient, compared to existing search methods, and it identifies subset candidates that merit attention according to some or all measures of joint influence that have appeared in the literature to date. Examples were given illustrating the method applied to two data sets previously analyzed in published studies.

Zhou et al. (2009) Proposed a three-stage k-means algorithm of $O(nkt)$ polynomial time was proposed to cluster the numerical data and detect the outliers. The clusters were preliminarily determined at the first stage. The local outliers of each cluster were found out and their influences on the centroid were removed at the second stage. Global outliers were consequently identified. Finally, the clusters, the densities of which were similar and some parts of which overlap, were merged. Simulation results show that the algorithm supports the discovery of clusters of different densities, different sizes and non-spherical shapes.
Pamula et al. (2011) proposed a clustering based method to identifying outliers. They apply k-means clustering algorithm to divide the data set into clusters. The points which were lying near the centroid of the cluster were not probable candidate for outlier and they can prune out such points from each cluster. Next they calculate a distance based outlier score for remaining points. The computations needed to calculate the outlier score reduces considerably due to the pruning of some points. Based on the outlier score they declare the top n points with the highest score as outliers. The experimental results using real data set demonstrate that even though the number of computations was less, the proposed method performs better than the existing method.

The existence of outlier always leads to inaccurate, even wrong results in data mining. Agglomerative hierarchical clustering was performed firstly, and then the outliers was identified unsupervisely from the top to down of the clustering tree. The computed results show that, the method can effectively detect global outliers, and the algorithm was efficient, user-friendly, and applicable to detect the outliers before data mining for high-dimensional and large databases. A proposed technique for detecting outliers while preserving privacy, using hierarchical clustering methods to analyze our technique to quantify the privacy preserved by this method and also prove that reverse engineering the perturbed data was extremely difficult. An effective and global outlier detection method was proposed by Jiang & bo An (2008), Liang (2010), Challagalla et al. (2010).

Chawla & Gionis (2013) present a unified approach for simultaneously clustering and discovering outliers in data. Our approach was formalized as a generalization of
the k-means problem. They prove that the problem was NP-hard and then present a practical polynomial time algorithm, which was guaranteed to converge to a local optimum. Furthermore they extend our approach to all distance measures that can be expressed in the form of a Bregman divergence. Experiments on synthetic and real datasets demonstrate the effectiveness of our approach and the utility of carrying out both clustering and outlier detection in a concurrent manner. In particular on the famous KDD cup network-intrusion dataset, they were able to increase the precision of the outlier detection task by nearly 100% compared to the classical nearest-neighbor approach.

Al-Zoubi et al. (2010) proposed a new efficient method was based on fuzzy clustering techniques. The c-means algorithm was first performed, then small clusters were determined and considered as outlier clusters. Other outliers were then determined based on computing differences between objective function values when points were temporarily removed from the data set. If a noticeable change occurred on the objective function values, the points were considered outliers. Test results were performed on different well-known data sets in the data mining literature. The results showed that the proposed method gave good results.

### 2.5.5 Distance-based methods

Distance-based methods were used to identify outliers based on the measure of full dimensional distance between a point and its nearest neighbors in a data set. Outliers were points that were distant from the neighbors in the data set. These methods generally define outliers based on a global view of the data set. Knorr & Ng (1998) introduced the notion of distance-based outliers, the $DB(p,d) – Outlier$
Definition 2.1. A data point $x$ in a given data set was a $\text{DB}(p,d) - \text{Outlier}$ if at least $p$ fraction of the data points in the data set lies more than $d$ distance away from $x$.

The parameters $p$ and $d$ were to be specified by a user. So different choices of $p$ and/or $d$ lead to different observations being declared outliers. The authors of this definition proposed also some efficient algorithms for finding distance-based outliers.

Angiulli et al. (2006) A distance-based outlier detection method that finds the top outliers in an unlabeled data set and provides a subset of it, called outlier detection solving set, that can be used to predict the outlierness of new unseen objects, was proposed. The solving set includes a sufficient number of points that permits the detection of the top outliers by considering only a subset of all the pairwise distances from the data set. The properties of the solving set were investigated, and algorithms for computing it, with subquadratic time requirements, were proposed. Experiments on synthetic and real data sets to evaluate the effectiveness of the approach were presented. A scaling analysis of the solving set size was performed, and the false positive rate, that was, the fraction of new objects misclassified as outliers using the solving set instead of the overall data set, was shown to be negligible. Finally, to investigate the accuracy in separating outliers from inliers, ROC analysis of the method was accomplished. Results obtained show that using the solving set instead of the data set guarantees a comparable quality of the prediction, but at a lower computational cost.

Y. Li & Kitagawa (2008) presented a method of outlier detection to identify exceptional objects that match user intentions in high dimensional datasets. Outlier detection
was a crucial element of many applications like financial analysis and fraud detection. Scholars have made numerous investigations, but the results show that current methods fail to directly discover outliers from high dimensional datasets due to the curse of dimensionality. Beyond that, many algorithms require several decisive parameters to be predefined. Such vital parameters were considerably difficult to determine without identifying datasets beforehand. To address these problems, it take an Example-Based approach and examine behaviors of projections of the outlier examples in a dataset. An example-based approach was promising, since users were probably able to provide a few outlier examples to suggest what they want to detect. An important point was that the method should be robust, even if user-provided examples include noises or inconsistencies. Our proposed method was based on the notion of DB- (Distance-Based) Outliers. Experiments demonstrate that our proposed method was effective and efficient on both synthetic and real datasets and could tolerate noise examples.

Sadik & Gruenwald (2010) Data stream was a newly emerging data model for applications like environment monitoring, Web click stream, network traffic monitoring, etc. It consists of an infinite sequence of data points accompanied with timestamp coming from external data source. Typically data sources were located onsite and very vulnerable to external attacks and natural calamities, thus outliers were very common in the datasets. Existing techniques for outlier detection were inadequate for data streams because of its metamorphic data distribution and uncertainty. They proposed an outlier detection technique, called Distance-Based Outline Detection for Data Streams (DBOD-DS) based on a novel continuously adaptive probability density
function that addresses all the new issues of data streams. Extensive experiments on a real dataset for meteorology applications show the supremacy of DBOD-DS over existing techniques in terms of accuracy.

Ramirez-Padron et al. (2010) Outlier detection was an important research topic that focuses on detecting abnormal information in data sets and processes. They addressed the problem of determining which class of kernels should be used in a geometric framework for nearest neighbor-based outlier detection. It introduces the class of similarity kernels and employs it within that framework. They also propose the use of isotropic stationary kernels for the case of normed input spaces. Two definitions of similarity scores using kernels were given: the k-NN kernel similarity score (kNNSS) and the summation kernel similarity score (SKSS). The preliminary experimental results comparing the performance of kNNSS and SKSS for outlier detection on four data sets. SKSS compared favorably to kNNSS.

Shaikh & Kitagawa (2012) Managing and mining uncertain data was becoming important with the increase in the use of devices responsible for generating uncertain data, for example sensors, RFIDs, etc. They extend the notion of distance-based outliers for uncertain data. To the best of our knowledge, this was the first work on distance-based outlier detection on uncertain data of Gaussian distribution. Since the distance function for Gaussian distributed objects was very costly to compute, they propose a cell-based approach to accelerate the computation. Experimental evaluations of both synthetic and real data demonstrate effectiveness of our proposed approach.
Angiulli et al. (2013) introduce a distributed method for detecting distance-based outliers in very large data sets. Our approach was based on the concept of outlier detection solving set Angiulli et al. (2006), which was a small subset of the data set that could be also employed for predicting novel outliers. The method exploits parallel computation in order to obtain vast time savings. Indeed, beyond preserving the correctness of the result, the proposed schema exhibits excellent performances. From the theoretical point of view, for common settings, the temporal cost of our algorithm was expected to be at least three orders of magnitude faster than the classical nested-loop like approach to detect outliers. Experimental results show that the algorithm was efficient and that its running time scales quite well for an increasing number of nodes. They discuss also a variant of the basic strategy which reduces the amount of data to be transferred in order to improve both the communication cost and the overall runtime. Importantly, the solving set computed by our approach in a distributed environment has the same quality as that produced by the corresponding centralized method.

2.5.6 Density-based methods

Density-based methods were proposed to take the local density into account when searching for outliers. These methods define outliers based on the local structure of the data set. Zhong & Huang (2012) have proposed density based outlier detection problem with low accuracy and high computation, a variance of distance and density (VDD) measure.
Momtaz et al. (2013) The problem of unsupervised outlier detection was challenging, especially when the structure of data was unknown. They presented a new density-based outlier detection technique that detects the top-n outliers. It overcomes the limitations of existing approaches, like low accuracy and high sensitivity to parameters. Our approach provides a score to each object called Dynamic-Window Outlier Factor (DWOF). DWOF improves Resolution-based Outlier Factor method (ROF) to consider varying-density clusters, which improves outliers’ ranking even when providing same outliers. Experiments show that DWOF’s average accuracy was better than existing approaches and less sensitive to its parameter.

Cao et al. (2014) Outlier detection was one of the key problems in the data mining area which could reveal rare phenomena and behaviors. It will examine the problem of density-based local outlier detection on uncertain data sets described by some discrete instances. They propose a new density-based local outlier concept based on uncertain data. In order to quickly detect outliers, an algorithm was proposed that did not require the unfolding of all possible worlds. The performance of our method was verified through a number of simulation experiments. The experimental results show that our method was an effective way to solve the problem of density-based local outlier detection on uncertain data.

2.6 Discordancy Test for Outlier Detection

The discordancy tests (single- as well as multiple outlier types) originally proposed by both W. J. Dixon (1951) and Grubbs (1950, 1969) have been very popular and were still in wide use in different fields (e.g., Barnett & Lewis (1978), X. Li et al. (2003),
Serbst et al. 2003, Farre et al. 2006, Gabrovská et al. 2006, Sang et al. 2006, Verma & Quiroz-Ruiz (2006), Hayes et al. (2007)). Several other discordancy tests were also available for this purpose, most notable among which were the skewness and kurtosis tests (Barnett & Lewis (1994), Verma (1997), Velasco & Verma (1998)).

Butler (1983) have motivated Outlier discordancy tests by considering the residual pattern recognition approach of Daniel and the approaches of cross-validation and maximum likelihood. Various optimality properties of the significance tests for the most outlying subset of data were shown and a multivariate two-way layout.

Giraudeau & Chastang (1999) have estimates of this coefficient could be influenced by outlying observations (the outlying mean or outlying variance of the measures). It was provided a procedure to detect these two types of outlier by means of approximate tests. Satterthwaite’s approximation was used to derive approximate probability density functions, and Bonferroni’s bound was then used to obtain two approximate tests allowing us to detect a potential outlier because of location slippage and a potential outlier because of dispersion slippage.

Verma & Quiroz-Ruiz (2006) have the modifications of the simulation procedure and accurate critical values or percentage points of nine discordancy tests, with 22 test variants, and each with seven significance levels for normal samples of sizes $n$ up to 100 were reported. Verma et al. (2009) has reported a geochemical data on reference materials (RMs) processed by outlier-based methods that use univariate discordancy tests.
Zijlstra et al. (2013) The sensitivity and the specificity of four outlier scores were studied for four different discordancy tests. The outlier scores were the Mahalanobis distance, a robust version of the Mahalanobis distance, and two measures tailored to discrete data, known as O+ and G+. The discordancy tests were Tukey’s fences (a.k.a. boxplot). Tukey’s fences with adjustment for skewness (adjusted boxplot), the generalized extreme studentized deviate (ESD), and the transformed ESD (ESD-T). Outlier scores O+ and G+ performed better than the Mahalanobis distance and its robust version. Discordancy tests ESD-T and adjusted boxplot were advocated for high specificity and ESD for high sensitivity.

2.7 Multivariate outlier

Detecting outliers in a multivariate point cloud was not trivial, especially when there were several outliers. The classical identification method did not always find them, because it was based on the sample mean and covariance matrix, which were themselves affected by the outliers. That was how the outliers get masked. To avoid the masking effect, they propose to compute distances based on very robust estimates of location and covariance. These robust distances were better suited to expose the outliers. In the case of regression data, the classical least squares approach masks outliers in a similar way. Also here, the outliers may be unmasked by using a highly robust regression method. Finally, a new display was proposed in which the robust regression residuals were plotted versus the robust distances. This plot classifies the data into regular observations, vertical outliers, good leverage points, and bad leverage points. Several examples were discussed.
Banerjee & Iglewicz (2007) A simple univariate outlier identification procedure was presented for the detection of multiple outliers in large and moderate sized data sets. This procedure was a modification of the well-known boxplot outlier-labeling rule. Critical values were easy to obtain for the large sample case for a variety of useful distributions, including the normal, t, gamma, and Weibull. Simple adjustment formulas and graphs were provided for handling smaller samples. Basic probability properties were obtained mathematically and through simulation. Two data sets illustrate the procedure’s application as a simple and effective screening tool for both moderate and large-sized univariate samples.

Dovoedo & Chakraborti (2015) described typical applications of boxplots include information about the underlying distribution as well as identifying possible outliers. This article focuses on a modification using a type of lower and upper fences similar in concept to those used in a traditional boxplot; however, instead of constructing the upper and lower fences using the upper and lower quartiles, respectively, and a multiple of the IQR, multiples of the upper and the lower semi-interquartile ranges (SIQR), respectively, measured from the sample median, were used. Any observation beyond the proposed fences was labeled a potential outlier. An exact expression for the probability that at least one sample observation was wrongly classified as an outlier, the so-called “some-outside rate per sample” (Hoaglin et al. (1986)), was derived for the family of location-scale distributions used in the determination of the fence constants. Tables for the fence constants were provided for a number of well-known location-scale distributions along with some illustrations with data; the performance of the outlier detection rule was explored in a simulation study.
2.7.1 **Robust Regression and Outlier Detection**

*Cook (1977)* proposed a new measure based on confidence ellipsoids for judging the contribution of each data point to the determination of the least squares estimate of the parameter vector in full rank linear regression models. The measure combines information from the studentized residuals and the variances of the residuals and predicted values presented with examples. *Draper & John (1981)* Statistics offered by *Cook (1977)* and Andrews and Pregibon (1978) purport to reveal influential observations in a regression analysis. Detailed examination of these statistics shows that two different types of influence were being measured and this was illustrated with examples derived from a set of data given by Mickey, Dunn, and Clark (1967).

*Walczak & Massart (1995)* proposed procedure for robust principal components regression based on the ellipsoidal multivariate trimming (MVT) and the least median of squares (LMS) methods used as an outlier detection tool. The performance of this approach was evaluated using simulated data randomly contaminated. *Pell (2000)* has described Robust statistical methods were less sensitive to outliers and could provide a powerful tool for the reliable detection of multiple outliers. They examines the use of robust principal component regression (PCR) and iteratively reweighted partial least squares (PLS) for multiple outlier detection in an infrared spectroscopic application. *Rousseeuw & Leroy (2005)* proposed several identification of outliers in regression analysis deals with unmasking outliers and leverage points in multivariate data using Mahalanobis distance (MD) and Robust distance (RD) along with diagonal elements of the Hat Matrix.
Gao et al. (2005) proposed a procedure for identifying multivariate outliers using Hawkins data set. Max-Eigen Difference (MED) method was briefly discussed about theoretical aspect of procedures to compare with Mahalanobis distance (MD) and Robust distance (RD) with examples.

Nedret & Gulsen (2008) The problem of outliers in statistical data has attracted many researchers for a long time. Consequently, numerous outlier detection methods have been proposed in the statistical literature. However, no consensus has emerged as to which method was uniformly better than the others or which one was recommended for use in practical situations. It perform an extensive comparative Monte Carlo simulation study to assess the performance of the multiple outlier detection methods that were either recently proposed or frequently cited in the outlier detection literature. Our simulation experiments include a wide variety of realistic and challenging regression scenarios. They give recommendations on which method was superior to others under what conditions. Dang & Serfling (2009) has introduced nonparametric multivariate outliers detection based on multivariate depth functions, also masking robustness against misidentification of outliers and non-outliers.

2.7.2 Outlier Detection in Time series Data

Fox (1972) and Abraham & Box (1979) discussed two models were considered for outliers and their effects in time series. Likelihood ratio and approximate likelihood ratio criteria were derived for these models and the power functions were compared with that of the approach generally applied in the past. Chernick et al. (1982) had investigate the effects of outliers using influence function for the estimator of the autocorrelation function (ACF) of a time series.
Abraham & Chuang (1989) had proposed some statistics used in regression analysis were considered for detection of outliers in time series. Approximations and asymptotic distributions of these statistics were considered. A method was proposed for distinguishing an observational outlier from an innovational one. A four-step procedure for modeling time series in the presence of outliers was also proposed, and an example was presented to illustrate the methodology.

Ljung (1993) had considered a estimation and detection of outliers in a time series generated by a Gaussian autoregressive moving average process. It was directly related to the estimation of missing or deleted observations shown that the estimation of additive outliers. A recursive procedure for estimates the Likelihood ratio and score criteria for detecting additive outliers were closely related to the leave-k-out diagnostics appropriate for innovational outliers.

Balke (1993) had demonstrates the difficulty that traditional outlier detection methods, such as that of Tsay, had in identifying level shifts in time series. Initializing the outlier/level-shift search with an estimated autoregressive moving average model lowers the power of the level-shift detection statistics. Furthermore, the rule employed by these methods for distinguishing between level shifts and innovation outliers did not work well in the presence of level shifts. A simple modification to Tsay's procedure was proposed that improves the ability to correctly identify level shifts. This modification was relatively easy to implement and appears to be quite effective in practice.
McQuarrie & Tsai (2003) proposed a new technique for detecting outliers in autoregressive models and identifying the type as either innovation or additive. This technique could be used without knowledge of the true model order, outlier location, or outlier type. Specifically, they perturb an observation to obtain the perturbation size that minimizes the resulting residual sum of squares (SSE). The reduction in the SSE yields outlier detection and identification measures. In addition, the perturbation size could be used to gauge the magnitude of the outlier. Monte Carlo studies and empirical examples were presented to illustrate the performance of the proposed method as well as the impact of outliers on model selection and parameter estimation. They also obtain robust estimators and model selection criteria, which were shown in simulation studies to perform well when large outliers occur.

Wu et al. (1993) Time series data often contain outliers which have an effect on parameter estimates and forecasts. Outliers in isolation have been well studied. However, in business and economic data, it was common to see unusually low observations followed by unusually high observations or vice versa. They model this behaviour by using a new type of multiple time period outlier which they call a reallocation, defined to be a block of unusually high and low values occurring in such a way that the sum of the observations within the block was the same as might have been expected for an undisturbed series. They derive tests for detecting reallocation outliers and distinguishing them from additive outliers. They show the effect on forecasts and forecast intervals of ignoring reallocation outliers. Finally, The apply our methods to two example data sets.
Changetal. (1988) Outliers in time series could be regarded as being generated by dynamic intervention models at unknown time points. Two special cases, innovational outlier (IO) and additive outlier (AO), were studied. The likelihood ratio criteria for testing the existence of outliers of both types, and the criteria for distinguishing between them were derived. An iterative procedure was proposed for detecting IO and AO in practice and for estimating the time series parameters in autoregressive integrated moving average models in the presence of outliers. The powers of the procedure in detecting outliers were investigated by simulation experiments. The performance of the proposed procedure for estimating the autoregressive coefficient of a simple AR(l) model compares favorably with robust estimation procedures proposed in the literature. Two real examples were presented.

Bilen & Huzurbazar (2002) considers the problem of detecting outliers in time series data and proposes a general detection method based on wavelets. Unlike other detection procedures found in the literature, our method did not require that a model be specified for the data. Also, used of our method were not restricted to data generated from ARIMA processes. The effectiveness of the proposed method was compared with existing outlier detection procedures. Comparisons based on various models, sample sizes, and parameter values illustrated the effectiveness of the proposed method.

Caroni & Karioti (2004) Tests for an innovative outlier affecting every member of a set of autoregressive time series at the same time point were developed. In one model, the outliers were represented as independent random effects; likelihood ratio tests were derived for this case and simulated critical values were tabulated. In a
second model, assuming that the size of the outlier was the same in each series, a standard regression framework could be used and correlations between the series were introduced. Simulation studies showed that approximate critical values obtained from the $\chi^2$ distribution work well for heteroscedastic independent series and for the case of equal correlations between each pair of series.

Shieh et al. (2006) a new robust fuzzy clustering approach was proposed for better performing principal component analysis (PCA) on function curves and character images that not only have loops, sharp corners, and intersections but also were bound of noise and outlier data. The proposed method was composed of two phases: firstly, input data were clustered using the proposed distance analysis to get good initial cluster centers and a reasonable number of clusters; secondly, the input data were further reclustered by the proposed robust fuzzy c-means (RFCM) based on the results obtained in the first phase to overcome the influence of noise and outlier data so that a good result of principal components could be found. Several function curves and Chinese character images were given to illustrate the effectiveness of the proposed method. Experimental results had demonstrated that the proposed approach works very well on PCA for both curves and images despite the fact that their input data sets may include loops, corners, intersections, noise, and outlier information.

Ferdousi & Maeda (2006) Fraud detection was of the great importance to financial institutions. It was concerned with the problem of finding outliers in time series financial data using Peer Group Analysis (PGA), which was an unsupervised technique for fraud detection. The objective of PGA was to characterize the expected pattern of behavior around the target sequence in terms of the behavior of similar objects,
and then to detect any difference in evolution between the expected pattern and the target. The tool has been applied to the stock market data, which had been collected from Bangladesh Stock Exchange to assess its performance in stock fraud detection. They observed PGA could detect those brokers who suddenly started selling the stock in a different way to other brokers to whom they were previously similar. They also applied t-statistics to find the deviations effectively.

Galeano et al. (2006) used projection pursuit methods to develop a procedure for detecting outliers in a multivariate time series. They showed that testing for outliers in some projection directions could be more powerful than testing the multivariate series directly. The optimal directions for detecting outliers were found by numerical optimization of the kurtosis coefficient of the projected series. They proposed an iterative procedure to detect and handle multiple outliers based on a univariate search in these optimal directions. In contrast with the existing methods, the proposed procedure could identify outliers without prespecifying a vector ARMA model for the data. The good performance of the proposed method was illustrated in a Monte Carlo study and in a real data analysis.