International Journal of Engineering & Technology, 7 (4.39) (2018) 189-198

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Outliers in Data Mining: Approaches and Detection

Deepti Mishra [1] *, Devpriya Soni [2]

[1]G L Bajaj Institute of Technology & Management, Gr.Noida, India,
[2]JIIT, Noida,
**Corresponding author E-mail: itsdeepti.s@gmail.com*

Abstract

The paper is grounded on the study of outliers which are the objects that somehow arise unlike from residue data stored and can be pointed as outliers. At present in data mining Outlier detection is the currently innovative topic for research. Outliers detection in a set of patterns is a pertinent problem in the data mining area. Outlier mining is the problem of detecting unseen events, abnormal data and exceptions. Another perspective of outliers they affect the outcomes and analysis of data. Presence of outliers make the results in confusable state. The patterns generated after the calculations from the data are not authentic and precise because of the outliers. This is the focus of this review as well as of that of this paper as well. There are some common categories of outliers described in this paper. In the residue of this paper, we will discuss briefly about data mining, outliers and their different categories, data mining techniques for outlier detection, application to support outlier detection from the data set, and approaches for outlier detection.

Keywords: *Data Mining; Knowledge discovery; Outliers; Outlier Detection; Overfitting*

1 Introduction

Outlier detection in an accumulation of patterns present in the data is a tedious problem in the data mining field. Outliers are the set of data points which acts differently from the residue set of data. We assume practically that the number of “normal” observations are considerably more than “abnormal” observations (outliers/anomalies) in the given data. Outlier mining is the problem of detecting deviant objects, unusual events, exceptionally different objects and exceptions from rest of the data set. Outliers are the data points those cannot be fitted in any type of clusters. These objects are somehow different from other objects in the data set. They may be different from complete data sets or may be difficult from its neighbourhood only. Till now, we can assume , there is no existence of standard methodologies for sensing outliers since it is very complicated to define features on which outliers can be marked out. As for now the outliers are categorized in various kinds, so it can be puzzling to notice all classes of outliers such as local or global. We can differentiate both outliers as, local will be abnormal from its neighbour only but global will be abnormal from complete data set.

Outlier detection, now a days the most researchable area in data mining field. As far as data mining is concerned the outlier detection is an vital subject for knowledge discovery.

The paper is comprising of literature survey for outliers and the approaches to detect outliers. There are many definitions of outliers proposed by researchers. Various techniques for outlier detection are described in the paper. Fundamentally, detection of outliers are the part of pre-processing of data in data mining techniques. Within the paper, we have explored the different methods of outlier detection in data mining. Initial easiest analytical step in outliers is to define the data – decoct its statistical variables (i.e. means and standard deviations SD). The identification and removal of noise and inconsistent data making data in presentable in more understandable format.

2. Knowledgebase for Outliers

2.1 Data Mining

Now a day there is abundance of data in the data base or data warehouses. currently Increment in Data is going on very rapidly. For such massive data, the Manual data analysis and data retrieval system is very time consuming [1]. The necessity of data analysis tools are required because of generation of enormous amount of data. This growing of data makes a realization for the use of data mining. A capable and robust tool can be constructed using data mining with its techniques. Data mining is an inspiring field of computer science. Scientists and researchers find it as a innovative field for their researches. Acquiring knowledge and decision making is a big aspect of data mining [2]. In other words we can say the data mining is the part of knowledge discovery [3]. Data mining is extraction of knowledge of from enormous databases [4]. It includes many techniques for extracting knowledge. It can handle large amount of databases. It is the procedure of refining the mysterious but beneficial knowledge from huge databases. Data mining techniques are applied to identify the patterns in the databases and image processing [5] , [6] . Pattern recognition is the emerging application of various algorithms and techniques for the purpose of recognition of patterns in the various kind of data. Data mining and statistics techniques can be applied to pattern recognition. Data mining is applied to analyze large data sets of genetic algorithmic data. Genetic algorithm is widely applied to data mining. Genetic algorithms are applied to improve the performance of data mining [7]. [8] data mining can be applied to detect anomalies in the various data sets. Numerous statistical techniques can be used to spot the anomalies. It is an significant issue of data mining [9]. Various data mining techniques can be applied in framework of education [10]. They provide a framework for education data mining. [11] Data mining enlightens a good understanding of the information and knowledge discovered from the resulted conclusions by applying its numerous procedures and practices that are beneficial to generate worthy patterns conclusions and patterns from the data set.

Data Mining is multidisciplinary topic of computer science and can be applied with such as artificial intelligence, statistics, machine learning and databases, neural networks, statistics [12]. Many a times, we apply statistical methodloges to identify the unusual patterns and abnormal data [13]. A division of mathematics which is known as Statistics is a applied to analyse and compute statistical calculation on data [14] . Various mathematical methods such as Bayesian Theory of probability, efficient estimation by maximum likelihood, principal component analysis, least squares and least absolute deviation estimation, Monte Carlo Markov Chain (MCMC) algorithm variance algorithm ANOVA, another MANOVA and further ANCOVA [15], [16]. Data mining is majorly helpful for the business end users in comparison to statisticians as they emphasis on statistics for their use. Data mining can effectively apply statistics techniques in its tools. So it is coupled with statistics and other major topics.

Usually, data mining can be applicable in the enterprise as either making a business intelligent solution using the products available in the market or developing data mining techniques [17]. By developing software of Data mining is a kind of analytical tools for studying and examining data. It permits users to analyse data from various ways such as in terms of variables or dimensions, classify it, and summarizing the identified relationships. Data mining is also applicable in social networks [18]. Social network data include web pages, interrelated data of people, places and things.

One of the application of data mining techniques is Web mining which is used to find information from data from internet which further contains documents on different sites etc. It can be divided in these domains as web content mining, web structure mining and web usage mining [19].

2.2 Data ware House

In current scenario as the data is increasing rapidly it is very difficult to manage it. Data ware house construction requires cleaning and integration of the data. It act as a pre-processing step for data mining by providing OLAP(Online analytical Processing) tools to ensure accurate and efficient examination of multidimensional data of different varieties which is must for effective data mining [20]. OLAP operations can be integrated with many data mining functions example classification, clustering, association prediction for better results.

Data warehouse is a data base maintain a discreetly from an organizations operational databases by consolidating historical data for analysis and integration shown in figure1. world wide defined definition is it should be first subject oriented and then integrated, thridly time variant and in the last satisfy the

property of non volatile for managements decision making process." [21].

Fig 1. Functioning of Data Warehouse

Thus, this is actually semantically reliable data repository further an implementation of decision support model [22] [23].

[24] The data warehouse system allows easy retrieval of information. It must shows consistent in the information. It is adaptable to change. It illustrates the information in timely manner. The information should be secured from unauthorized access. The data warehouse provides good decision making regarding the data.

2.3 Knowledge Discovery in Databases

The acronym of Knowledge Discovery in Databases is KDD. The key factor in datamining is to extricate or mine the knowledge from huge databases. In other words knowledge discovery can be stated as knowledge mining [4] [25]. Knowledge discovery and data mining share the same aim that is to discover useful data from large databases.

Fundamentally the resulted patterns generated after applying data mining techniques to find some conclusions is termed as knowledge discovery.

We can say, knowledge discovery is the process of finding exact and accurate conclusions from data, and have a significant work which is generated by data mining Data plays a vital role as a tool to gain a competitive edge by supporting improved and customized services. Knowledge discovery is the process which consists an iterative sequence of following steps : -

The initial step of the KDD process is to identify the goal for KDD process from the users view point. After setting the goal user select the data set or subset of data set on which KDD process is to performed. While selecting the data set user needs to ponder in right direction.

Then organising and cleaning of data , data amalgamation, choice of data, data alteration, following the steps of data mining(essential step), conclusion generation steps are to be followed.

3. Outliers

Outlier Detection is a key work in many crucial applications and data sets such as in the conditions like aircraft defect, Fraud detection in banks [26].Outliers are the critical problems in various disciplines of knowledge data bases such as data mining, statistics , artificial intelligence and machine learning [27]. Outlier detection can be applied for intrusion detection, frauds in medical cases, novelty detection, intrusion detection [28]. Outlier detection process is used in various applications example fraud detection in banks and financial sectors, intrusion detection system, tracking environmental activities, healthcare diagnosis [29] & [30]. Outliers are also big troubling topic in the linear models as the Anova techniques of statistical methods shows prone to such deviating data points [31].

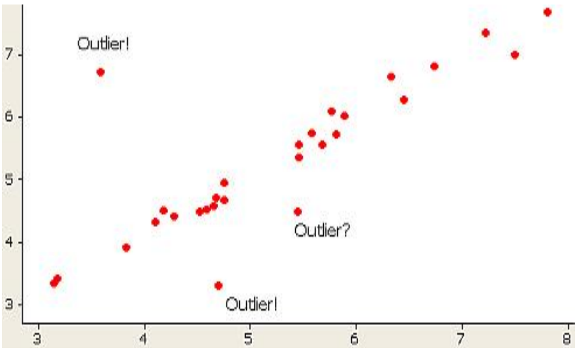

Fig 2: Outliers

3.1 Definitions of Outliers by Different Authors

1. Hawkins [32] – An outlier as an spotted data point, which diverges within the data set with other data points such that creating a doubt that it was intuitive by dissimilar procdure.
2. Barnett and lewis [33] – An outlier which is different from the residue of the data set.
3. [34] An Outlier is the point that can be categorized with other data points normally.
4. [35] The data object which is counted as an outlier, can cover important data with respect to the data set.
5. Outlier is a pattern or a data point or small quantity of data object those are alike with residue of the pattern in the data

set based on some measure from the general distribution of the data. [36] [37] [38] [39] [40] [41] [27]
6. [42] Outliers can be generated due to wrong observations in the experimental results.
7. Breuning et al. – Density is the basic characteristics to spot the outliers, as they lie in lower density region.
8. [43] provides the definition for depth based outliers as the points in the shallow convex hull layers.
9. [44] Outliers are those objects that may not be applied to any clusters of the data set
10. [45] provides the definition of outliers for graph based methodology as for data objects those are present in the particular position of the graph that makes them outliers.
11. [46] The spotting of outliers can follow the approach as defining a point if its density value varies wit other data points in the region whether may be high or low.
12. [47] In the feature space, outliers can be detected as the points that are distanced from other points in the data set.
13. [48] That point can be considered as outlier, if its deletion from the set makes the result more precise and accurate.
14. [49] For detecting spatial outlier, spatial neighbourhood is taken into consideration as the point is significantly different.
15. [50] Outliers in spatial and temporal data sets, are detected as they have different characteristics and features from their neighbourhood.
16. [51] outliers are the minor subset of dataset so that the removal of that dataset create the resultant dataset in more precise state.
17. [52] Outliers are the observations those have different distributions from rest of the data set.
18. [53] defines outliers in vector space which referred as feature space as those points which lie in the sparse region of the feature space.
19. [54] outliers are those points in statistics those appear very different from entire data set.
20. [55] stated outliers as minority data points in the entire data set.

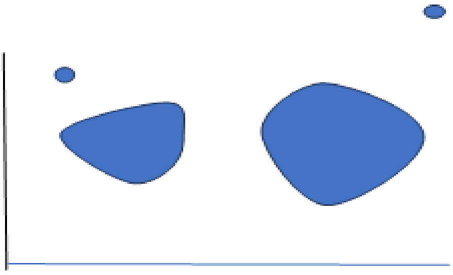

Fig 3: Pictorial representation of Outliers

3.2 Causes of Outliers

[56] defines the causes of Outliers in the data set.
1. Malicious activity – example frauds in hacking or criminal activity such as credit cards values in insurance data.
2. Instrumentational fault – for example faults in the machines or system, or wiring and cables.
3. Transformation in the environment – for example change in the buying patterns of the customers, change in the climate.
4. Human error – for example typing mistake, a data reporting error

3.3 Reasons for handling Outliers

[57] defines the reasons for handling the outliers.
1. Outliers have a significant impact on the results of the databases.
2. Outliers are considered as recorded errors, but some of them can be interested and useful for the conclusions and results.
3. In many research areas, these outliers can be the key in the discovery of unpredicted knowledge.

[58] considered as outliers as the key branch of data mining which has many application areas and requires more consideration.
[59] observes as outliers frequently share similar statistical characteristics, so it can be tough to differentiate between them. If there is lack of appropriate domain knowledge which drags to the concept of deciding that why they are outlying and what is the mechanism of the underlying data generation is then it is difficult to pointing them as outliers. To detect whether an outlier is revealing or valuable some other information is required example precise mathematical calculations, appropriate domain etc. [60] Outlier detection algorithms have objective to automatically spot those objects or troubling observations in huge amount of data. As there is not any typical definition for outlier, so every algorithm is based on a model that is relying on certain assumptions of what qualifies as an outlier

4. Applications of Outlier Detection

The requirement of accuracy, the detection of outliers are the significant fragment of data computation. Following are a few real life applications that have been widely applied for numerous application domains [27], [26]. The list I believe is not exhaustive.
1. Detection of fraudulent- detection of fraud values in credit cards, debit cards, use of mobile phones or insurance claims, financial transactions etc.
2. Intrusion detection- Outlier detection helps out to detect unauthorized access during communication by restricting hacking.
3. Functioning of Network- Nursing and checking of computer networks to find blockages.
4. While checking Loan application for further processing to detect fraud applications or potentially problematic customers.
5. Activity monitoring- Outlier detection can be applied to detect fraud in the quity market by finding the unauthorized phone activities.
6. Environmental monitoring example Cyclone , Tsunami , Floods, Drought, Fire, Typhoon by detecting a typical environmental behaviour to minimise losses of life and material.
7. Public health- for detecting unusual symptoms or abnormal test results to identify instrumentation / clerical errors or real health problems.
8. Medical conditioning monitoring example heart rate monitors
9. Pharmaceutical research- recognizing novel molecular structure.
10. Localization and tracking – Outlier detection can be applicable for finding and tracking of location of moving objects by filtering the data.
11. Logistics and transportations- The detection of outliers in shipments can be beneficial as it tracks the packages and shipment and try to maintain quality of work.
12. Noticing unforeseen entries in databases – Detection of Outliers are used in data mining to spot faults and extraneous entries.
13. Detecting wrongly labelled data in a training data module.

5. Characteristics of Outliers

Depends on multiple aspects for example location of outlier concerned, type of data, size of data, features taken into consideration and techniques utilized for identifying outlier. Below mentioned are the characteristics of outliers in the same fashion [27].

1. Types of detected outliers- Outliers are noticed grounded on their features and characteristics. The can be categorized in global and local outliers. We can differentiate both outliers as local will be abnormal from its neighbour only but global will be abnormal from complete data set.
2. Different Degree for considering outlier- a **SCALAR** (**binary**) manner to find that the data object is actually an outlier or not. In contrast **OUTLIERNESS** styles visualize the characteristics and degree on which the point can be considered as an outlier in respect with other data. It can be counted as outlier score.
3. Distinguishing outliers on basis of dimensions- A **UNIVARIATE** data which can be classified as, which has a sole attributes can be distinguished is an outlier only on the basis that a single attribute is an strange with respect to that of the other data. In contrast a **MULTIVARIATE** data which contains numerous attributes, further can be considers as an outlier as few attributes together have irregular and abnormal data values
4. Number of outliers detected at once- The count of outliers can be detected by various techniques in different ways in a moment. A few techniques detect one outlier and remove it and then repeat the procedure till no further outliers are detected. Some techniques detect a group of outlies in single attempt. Both have their pros and cons.

6. Relation between Outliers and Data Mining

Data mining techniques and automated tools are required to assist the research of scientists and researchers and experts of application domain. As the data is increasing day by day it is observed that data mining tools are requisite to utilize data for marketing, commercial and scientific purpose.

Previously the researchers and scientists are managing small scale data having less number of attributes and less number of records still not finding exact results. But now as the size and dimension of data is increased a lot , it is very difficult to discover precise and accurate results and patterns . [28] stated that advancement in computers and internet has led to amassing of large amount of information that needs to be processed for each outlier detection task. This mandates an effective and efficient outlier detection methodologies on large scale. Free text is the commonest recorded human interaction on net (eg mail, blocks etc).so now, a challenging problem is to detect outliers in hefty text document which is important for finding fascinating items or suspicious documents which are odd man out. Since the size of the data is increased rapidly, so there is also increment in the unwanted data. That unwanted data is named as outliers [61]. That unwanted data has to be detected and removed from the original data. Some unwanted data is good for analysis which concludes towards knowledge some are bad outliers that deviate the results with their presence [62]. Presence of outliers will affect the result concluded for knowledge discovery. Sometimes outliers will provide the meaningful information about the result and sometime they affect the results badly. This thesis present the outlier detection algorithm in bivariate data set.

7. Types of Datasets Used for Outlier Detection

[27] discussed that multiple outlier detection methods worked differently for different sets of data types. Characteristics and attributes of data of the virtues that divide data sets into simple and complex types. Complex type is further is sub classified into six sub types depending upon different semantics of data as mentioned below. It is the complex data set that is a matter of major concerned for the outlier detection problem.

1. Simple data set – it refers to frequently applied data set having low dimensional real valued ordering attributes with no complexities Generally simple data set is easy to handle and can easily applicable to the algorithms and techniques.
2. High dimensional data set - the high dimensional data is the representation multi dimensional objects [63]. Some examples can be Geographic information system (GIS), Pattern recognition database (PR), Image Processing databases (DIP), where the data is made up of a set of objects with various number of attributes. These attributes can be defined by feature vectors in multi dimensional space.
3. Mixed type Attributes – It contains mixture of continuous (numeric) and categorical (non- numeric & partial ordering) values.
4. Sequence data set- as some kind of data sets are designed and formed as a arrangement of distinct entities eg symbols or letters and therefore the data does not has same length and no priori distribution.
5. Spatial data set- It has spatial (eg location ,shape, directions , geometric , topographical info.) and non- spatial (intrinsic info of data) attributes.
6. Streaming data set- it can be think like that a huge amount of is coming very fast and unbreakable fashion seen in many real time applications and unlimited in size.
7. Spatio temporal data set-it refers to the temporal and spatio temporal relationships existing among spatial data as a attributes usually seen in many geographic phenomena evolving over time (eg traffic analysis, mobile computing).

8. Types of Outliers

Outliers can be organised on the basis of their characteristics and concepts [64] , [65] & [66] & [67].

• Point outliers. In a set of data set, a different data object can be considered as a point outlier. They can be measured as the data points that differ on their own in comparison to other data points in the set. The point outliers can be detected as counting and measuring their deviations from other data points in the set.

• Collective outliers. It is somehow different from point outliers as the number of points that are considered as outliers are the group of some data objects. The complete group of data objects those deviate from rest of data objects can be considered as collective outliers. Generally the collective outliers have some common characteristics.

In collective outliers one more term can be added known as sequence outliers. When the data is represented in the form of sequence, the collection of outliers in that sequential data are termed as sequence outliers. It may be log entries, buffer entries etc.

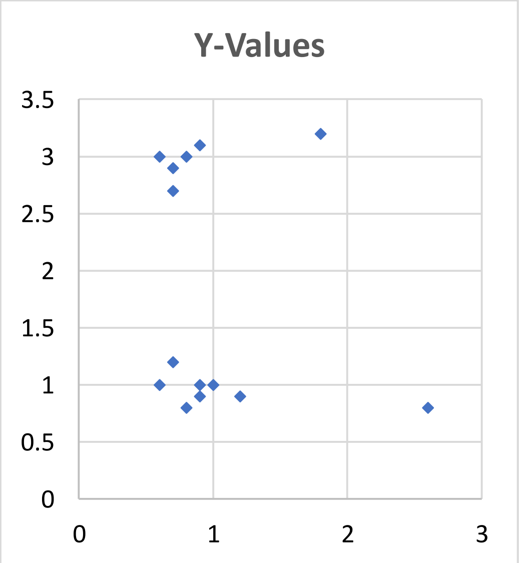

Fig 4: Showing clustering and outliers

• Vector Outliers. Generally when the dimensions and properties are increases of the data set it can be represented in the form of vector like representation. The outliers that usually detected from such kind of databases can be termed as vector outliers.

• Trajectory outliers. As like the term, the outliers those are related to moving objects , or the paths related to moving objects, changing in vehicle positions, tracking data of storm etc can be considered as trajectory outliers. They may be point outliers or collective outliers.

• Graph outliers. When the data is represented in the form of graph that is edges, nodes than the points deviating from the graphs are labelled as graph outliers.

A global outlier has a visible and remarkable data value which van be differentiated from rest of the data. For example, if 149 out of 150 data points have values between 800 and 1000, but the one point has a value of 5, than that point may be a considered as global outlier.

The definition of a local outlier is a calculation of model object that has a value within the standard and regular range for the whole dataset of objects, but if you look at the surrounding points, it is unusually high or low.

9. Problem of Overfitting

The question of overfitting depends on the preciseness of the model whose estimation is based on the sample and trained data set, can be optimistic estimation if the expected model inclines to overfit the data [68]. The information in the data has two origins: the first is what you like: patterns, relationships, between variables that exists in your real world. The user want to catch those patterns into the model. The second point is user do not like the noise, random, variation, sampling errors, unpredictable values,. Overfitting occurs when the modelling algorithm includes information of the second kind into the model. Categorical variables leads to overfitting to much more complex level. But categorial values do not typically have outliers [69]. Their values cannot be compared in terms of greater than or smaller than. More division of categorical values means more values the user have which incorporate more noise to the model. The solution to the problem is first to convert them to continues variables, second diminishing the number of values.

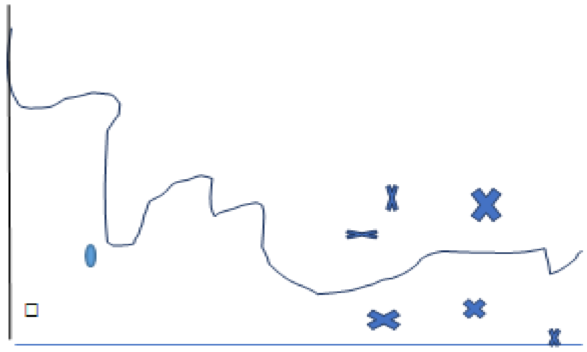

Fig 5: overfitting

10. Different Types of Approaches for Outlier Detection

A lot of research has been presented already for detecting various outliers from different kinds of data.

According to [4] there are three basic techniques for outlier detection. These approaches are statistical approach, distance based approach and deviation based approach.

Statistical Approach

In the approach a probability model is supposed and prepared for outlier detection to detect outliers. The prerequisite is the information of some parameters such as data distribution, values of mean and variance, targeted value of outliers.

A statistical discordancy test compares 2 hypothesis: - one is considered as working hypothesis and other is termed as alternative hypothesis for set of n objects in a distribution for considering them as outliers. One proved that they come from they are outliers and other states that they are the part of other distribution, in case of failure of first hypothesis.

Outlier discordancy tests are comprises of the residual pattern recognition approach of Daniel and the approaches of cross-validation and maximum likelihood [31]. Generally discordancy test is conducted in statistical based approach.

The pitfalls of this approach is that it can handle single attribute at a time, and many data mining algorithms needs multidimensional approach. It also needs prerequisite of the parameters for the data set, eg data distribution. May be possible the result is not precise.

10.1 Distance Based Approach

The distance based approach assumes those points as outliers who do not have adequate neighbours. And the neighbours are defined on the basis of calculated distance.

The technique avoids the cumbersome calculations related to observed distributions. It requires parameter values. Distance based approaches are applied generally on low dimensional data set to detect outliers.

Various procedures of distance based techniques are developed such as Index Based algorithms, Nested loop algorithm, cell based algorithm.

The first procedure applies R trees or KD trees which are the part of multi dimensional indexing structures calculates the radius value for each data point for its neighbours.

The second procedure deals with the memory by distributing buffer into two halves additionally separating the data into logical blocks.

The third procedure is applied on memory resident data set.

10.2 Deviation Based Approach

The key criteria of Deviation based approach is to identify data points based on their qualities and features within the group. Some of the data objects which deviates from the characteristics are well-thought-out as outliers.

Procedures for this approach are sequential exception technique and OLAP data cube technique.

The first procedure discriminates unfamiliar items from a series of data set. The second procedure is applied in large multi dimensional data to detect anomalies by applying data cubes.

[70] Further describes and classifies the methods of outlier detection in univariate, multivariate techniques and parametric and non-parametric techniques. These techniques include the above mentioned methods for outlier detection example statistical method is included in parametric technique, non parametric technique comprise of distance based methods. Univariate outliers can be detected using Box Plot and multivariate outlier can be detect using Mahalanobis distance calculation [71].

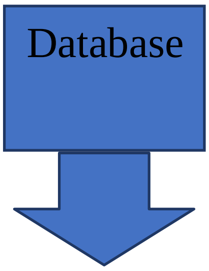

Fig 6: Pictorial representation of outlier analysis

The author Jingke Xi defined the outliers as the objects those deviate too much from other objects in the data set such as they create a doubtful result. [2] It creates an assumption that they were intuitive and presented through a different procedure as from other objects in the data set. The paper discusses and compares approaches of different outlier detection techniques from data mining perspective. According to the paper there are many applications of outlier detection such as ecology and many more. As per the paper there are two categories of outlier detection one is classical outlier approach and other is spatial outlier approach. As discussed in the paper, the previous approach further classified into statistical based approach, distance based approach, deviation based approach, density based approach. The classical approach is based on transaction data set which contains collection of items. A good example of transactional database is market analysis. Statistical based approach can be applied, for probability and distribution models for the given data set that fits the conditions to detect outliers. But this approach does not function well when the dimension increases. Distance based approach based on calculations of k nearest neighbours. Deviation based approach considers the characteristics of the data points and the point that deviate too much considered as outliers. Density based approach assumes those points as outliers those lies in density region.

A five step procedure was introduced for outlier detection [62]. The data cleaning techniques applied on the data set after that outlier detection is performed. The paper defines the four types of Data set – nominal, ordinal, interval and ratio scaled. There are Three types of Outlier detection techniques based on supervised learning, semi supervised learning and unsupervised learning. There exists three types of data univariate, bivariate and multivalued data set. Five types of approaches defined by the paper are statistical tests, depth based tests, distance based tests, density based test and cluster based test.

Reducing outliers are the mandatory task for achieving knowledgeable outcome [72]. The paper outline the major tasks in the data mining – classification, regression, clustering, summarization, dependency, change and deviation.

Outliers are the data objects that does not obey to the standard data objects illustrating the data set [34]. Detection of outliers is defined as the part of data cleaning. The paper presents the new algorithm PLDOF and tried to proof that it is better than LDOF. K means clustering algorithm is applied to divide the data set into clusters. It is based on the concept that first calculates the distance of each point from the centroid of the cluster. Now, If the calculated distance is smaller than the radius of the cluster, then that point is cropped. To compute the outlier points from the rest of the points in all the clusters.

Outliers are defined as one of the subtopic of data mining [35]. The paper defines the new technique ODHDP named (Outlier detection in high dimension based on projection). The procedure finds the outliers which is further grounded on projection technique, from the data set. It defines the outlier as the background noise in the data set. In first step clustering the projection of data set in each dimension, in second step dimension has maximum weight are selected. And in third step those clusters are pruned having less number of data values less than threshold value. In the last step outliers are detected.

Some authors also used distance based methodologies for detection of outliers. A new hybrid approach has been introduced using distance based and k means method to identify outliers [36]. The approach apply clustering algorithm that is k means for partitioning the data set into number of clusters and then apply distance based approaches to detect outliers. The paper defines the outlier as a pattern which is dissimilar with respect to the residue patterns. There are the different types of the techniques for detecting outliers- model based, connectedness, density based, distance based, cluster based, k nearest neighybour based.

The authors published a paper describing the outliers in high dimensional data set Charu c Agarwal [37] The paper is focussed mainly on outlier detection and its techniques over high dimensional data set. it defines the outliers as the background noise. The paper outlines properties that should assure for high dimensional data during outlier detection. The paper discussed a new technique - some evolutionary algorithm for outlier detection in high dimensional data set. the algorithm detects outliers by discovering the density distribution of projections from the data.

The technique identify the outliers by inspecting those projections of the data which have abnormally low density. The algorithm designed for outlier detection in the paper shows that it will handle the problems of high dimension.

The author suggested a new algorithm for outlier detection which is based on evolutionary search [73]. The genetic algorithm is applied to detect outliers. The methodology used is union of distance based Euclidean method, density based component which is further based on Lancaster's modified mean value. Author defined the outlier detection in data, and stated that they can be considered into three groups – statistical method, deviation method and distance based method.

Outliers can be defined as small quantity of the data objects with abnormal behaviour [38]. There are many applications of data mining such as business, weather forecasting, bioinformatics records. The paper introduced a new algorithm SWHOT which is based on weighted hypergraph model. It is the union of BSWH algorithm and CURE algorithm. The provide the concept of the feature vector and attribute similarity.

As suggested by authors, there are different groups of data objects can be stated as outliers [74]. The experiments are conducted on high dimensional data set for outlier detection algorithm. A new algorithm Angle based outlier detection (ABOD) is proposed and compared with other algorithms. The key gain of the approach is that to achieve the ranking, the mentioned method does not count on any parameter selection process. Unrelated features and characteristics are likely to occur in HDD. Algorithm considers the variances of the angles for calculation between the difference vectors of data objects. Distance also taken into account in the provided method. ABOD provided a better ranking w.r.t. the precision of the top ranked objects.

Authors introduced another approach for outlier detection for high dimensional data set [39]. This approach is effective and applicable for high dimensional data set. The functioning of procedure associates the methods named as random hyper plane projections and AMS sketch on product domains. It also enables to lessen the computational complexity. The approach is applicable in parallel environment to achieve parallel speed up.

The authors presents a new approach for outlier detection in high dimensional data set [75]. Authors did some modifications in Mahalanobis distance function and proposed a high break down minimum diagonal product estimator. By applying this concept, an algorithm for outlier detection is developed.

The author proposed a new framework that comprises of two processes. [76] First one is longer algorithm called PCOut and another method is named as Sign. It is based on Principal Component Analysis. The algorithm detects the location of outliers and then scattered outliers. The algorithm is based om the calculation of covariance matrix, eigenvalues and vector calculations. But algorithm retains only those eigenvalues that contributes maximum to covariance. According to author the algorithm can be extended and applied to genetics.

Outliers can be identified as anomalous objects [42]. In the paper a new algorithm is proposed to detect outliers named outlier finding technique (OFT) by doing clustering of data which practices k means algorithm. The approach functions on density grounded and another distance grounded methodologies for finding outliers.

Authors proposed a new algorithm for outlier detection [77] . In first step , authors modify the K-means clustering algorithm and then in step 2 constructed a MST(Minimum Spanning Tree). In step 1 a new cluster center is assigned to a new data pattern. And in step 2 it will remove the longest edge while creating MST. And in the last the small clusters having les number of nodes are selected and considered as outlier.

Outliers can be exceptional or rare cases from the data set [78]. In the paper a novel definition of outlier that is class outlier is proposed with defining a related term class outlier factor (COF). It provides and applied to determine a level of presence of a Class outlier for an object. The significant steps of processing COF are, initially calculating the probability of occurrence of the classes, the deviation within the same class following the distance among the instance with its existing neighbours.

Outliers creates many problems in the applications of data mining [40]. So outlier detection is a mandatory part of data mining. Applications of outlier detection are fraud detection, financial analysis, network robustness analysis. These kind of application generally comprises of high dimensional data set. A new approach is proposed in the paper which is combination of Subspace based and Example based methods to detect outliers in the data set.

Distinguishing anomalous data from the data set is very indispensable task [79]. It becomes more critical if the data is high dimensional. A new approach COMPREX is introduced which apply pattern based compression. The approach is parameter free, general, scalable and efficient. The categorical database is used.

The author presented a wavelet based approach to identify outliers. It operates on image data by applying wavelet transformation technique to determine regions having spatial variation. [80] Wavelet transformation is applied to the input data set which further generates wavelet power values from the input data. The calculated values are stored which are greater than the provided values of wavelet power in the algorithm. The suspected region which can contain outliers is reversed back to the original data by applying inverse of wavelet transformation. The suspected region can be stated as outliers. Authors focus on two dimensional region for outliers.

A new algorithm Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is applied to generate clusters [81]. Identifying patterns from the large data bases is a critical task. The algorithm operates on large databases. It can also tackle noise effectively. The paper defines the two types of attribute – metric and nonmetric. It functions in four phases – Load database in memory after building CF tree, reduce in smaller CF tree, apply global clustering, refine the clusters.

The authors suggested a new algorithm RBRP for detecting outliers. [82] It is fast algorithm which is based on distance based approach. It particularly function on high dimensional data set.

A new technique for clustering has been presented for large databases [83]. The principal idea is to conduct clustering in two steps: a rough and fast step such that the data is divided into overlapping subsets named as “canopies”, and then distance measure is done only for points that occur in common canopy. The canopies are the subset of the of the data points. The clustering based on canopies can be suitable to many existing clustering algorithms for example K means, greedy agglomerative clustering and expectation maximization. In the paper the canopies are applied with agglomerative clustering.

Authors proposed a new method for outlier detection. [84] It is based on the distance based approach for outlier detection where outlier are detected on the basis of distance of a data point from its k nearest neighbour. A rank is allotted to the data point on the basis of its k nearest neighbour.

Algorithm presented a outlier detection method for large datasets. [85] The algorithm functions on LCF(Local Connective Factor) which can identify both outliers and clusters in the meantime. Authors also introduced a vertical data representation named P-trees.

Authors proposed a outlier detection method which is based on distance based method for uncertain data of Gaussian Distribution. [86] They proposed a cell based approach as the distance function for Gaussian distribution is costly to calculate.

The spotted anomalies in the multidimensional arrays while each dimension represents categorical data. [87] Author detected anomalies by comparing current data with historical data. The methodology is based on empirical Bayes method further functions using two component Gaussian mixture to check deviations of data at current time with historical one.

The authors presented a methods for outlier detection for categorical data. [88] It is based on Hypergraph model in which each vertex denotes data object in given data where every hyperedge shows collection of objects those containing frequent item set. It functions in three steps as first one is building the hierarchy of the hyperedges, second step is constructing the multidimensional array and third step is detecting outliers in the array.

Authors discussed various methods to detect outliers in categorical data set given by different authors such as AVF (Attribute Value Frequency), NAVF (Normally distributed Value Frequency) [89] [90], ROAD algorithm , ROCK algorithm [91].

The author investigate the process of outlier detection for text documents. [28] The process become more complicated as the increase in the amount of collection in text documents. They applies the samples of the Fourier and Cosine spectrum for performing clustering on text documents. The generated outputs shows that Fourier Transform provides better and precise results in comparison with random projections.

The author implemented the algorithm for outlier detection named LODI. [92] The algorithm detects outliers which are maximally separated from its neighbours by presenting the quadratic entropy to adaptively select a set of neighbouring occurrences, and a learning process to pursue an optimal subspace.

The authors perform the clustering techniques in data streams and spot the outliers in data streams [93]. They identify outliers in data streams, the technique is the combination of two clustering algorithms namely BIRCH with K-Means and Birch with CLARANS.

Authors proposes a new algorithm for clustering stream data named D-Stream using a density based approach. [94] The algorithm uses two components one is online module and other is offline module. Online module maps each input record into a grid and offline module calculates a grid density and identify the clusters. The algorithm calculates clusters very efficiently and effectively. It applies a density decaying technique and attraction based mechanics.

The author Proposes an algorithm which functions on disk resident data sets and where I/O costs agrees to the cost of consecutively reading the I/P dataset file twice. [95] The proposed algorithm is modest to execute and can be applied with any type of database. Further More a change in the key method lets Dolphin to manage with the state where existing buffer of main memory is lesser than the standard requires. It attains proficiency by naturally integration with a unified schema and its strategies. The cost of I/O of the algorithm is very little as it matches to the cost of successively reading the I/P database file double. According to the author the methodology is very efficient to manage huge disk resident gathered data.

Author defines a sampling algorithm which detects distance-based outliers in domains. [96] The algorithm detects kth neighbourhood outlier. The distances from neighbourhood can be calculated on a random sample of data instead of whole data set.

The author proposed a framework which is applicable for stream data for clustering and named the approach as D-Stream, which is further based on density-based technique [97]. This procedure functions in two steps which includes two components one is online and another is offline, as in first step online component plots each input data record into grid. In second step there is an offline component which computes the grid density and clusters the grids based on the density.

Authors define that the work of outlier detection as main objective as detecting outliers to expand the analyses of data and eventually finding fascinating and advantageous facts about unseen events within various application areas. [27] The authors provide the description on current unsupervised outlier detection methods for numerous kinds of databases and offer a complete taxonomy outline for outlier detection techniques. They provide the description of the various outlier detection techniques for the users to select suitable method for datasets.

Authors presented an algorithm for discovering distance based outliers and illustrate that it can achieve near linear time scaling which applied with real time data. [98]

Authors provide a data distribution approximation framework which does not need a priori knowledge of the input and outline an efficient method for distribution deviation detection in sensor data. [99] The paper show that the framework can be extended to detect density based or distance based outliers.

Authors presented a new algorithm for outlier detection. [100] It calculates spatial local outlier measure(SLOM) value of an object o. Object with its SLOM value is compared with its neighbourhood. The variance of the neighbourhood is also captured as the neighbourhood plays the vital role in calculating outliers in spatial data.

Author proposed a new algorithm for outlier detection which is combination of fuzzy set theory and kernel functions. [101]

The author presents a concept of conceptual learning named COBWEB which is applied for hierarchical conceptual clustering using an incremental approach. [102] The system applies a hill-climbing search using a space of hierarchical classification schemes which further uses operators that allow bidirectional travel by this space. The authors define that the assumption

11. Conclusion

However, data mining has come a prolonged journey since the term came in existence. The suggestions of Data mining pledges in supporting organizations to reveal patterns veiled in their data such that, they can be applied to expose some applications like the behaviour of customer, products and developments. But there are many issues in data mining those need study and research. And outliers is one of them. They may be different from complete data sets or may be difficult from its neighbourhood only. The work presents the review of outlier and outliers detection techniques. The study comprises of analysis of various outliers and techniques used for outlier detection. Many authors outline different definitions of outliers per their work and study. For example outliers are defined in vector spaces, for high dimensional data sets, for density based clusters etc. There are numerous methodologies for finding outliers in different types of data sets. According to the type of application and data sets, it is to be decided by the programmer which outlier detection technique is the most suitable and beneficial for the application. Some techniques are designed for low dimensional data set, some algorithms are designed for vector space, some are based on density based approaches ,some are based on distance based approaches etc.

References

[1] M. R. Anderberg, Cluster Analysis for Applications, Academic Press, 1973.

[2] B. Dawson and R. G. Trapp, Basic and clinical Biostatics, Mc Graw Hill, 2004.

[3] J.P.Baride, A. Kulkarni and R. Mazumdar, Manual of Biostatics, Jaypee, 2003.

[4] K. H. Tung, J. Han, L. V. S. Lakshmanan and R. T. Ng, "Constraint Based Clustering in Large Databases".

[5] P. K. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, 2006.

[6] Atkinson, Exploring Multivariate Data with the Forward Search, New York: Springer-Verlag, 2004.

[7] M. Steinbach, L. Ertoz and V. Kumar, "The Challenges of Clustering High Dimensional Data," pp. 1-33.

[8] D. H. FIsher, "Knowledge Acquisition Via Incremental Conceptual Clustering," Machine Learning 2, pp. 139-172, 1987.

[9] M. Petrovsky, "Outlier Detection Algorithms in Data Mining Systems," Programming and Computer Software, pp. 228-237, 2003.

[10] P. Sun and S. Chawla, "On Local Spatial Outliers," in ICDM, 2004.

[11] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki and D. Gunopulos, "OnlineOutlierDetectioninSensorDataUsing Non-ParametricModels," in ACM VLDB, 2006.

[12] S. D. Bay and M. Schwabacher, "Near Linear Time Detection of Distance-Based Outliers and".

[13] Y. Chen and L. Tu, "Density-Based Clustering for Real-Time Stream Data," in ACM SIGKDD, 2007.

[14] M. Wu and C. Jermaine, "Outlier Detection by Sampling with Accuracy Guarantees," in ACM KDD, 2006.

[15] F. Anguilli and F. Fassetti, "DOLPHIN: An Efficient Algorithm for Mining Distance -Based Outliers in Very Large Datasets," ACM Transactions on Knowledge Discovery from Data, 2009.

[16] L. Tu and Y. Chen, "Stream Data Clustering Based on Grid Density and Attraction," ACM Transaction on Computational Logic, pp. 1-26, 2008.

[17] S.Vijayarani and P.Jothi, "An Efficient Clustering Algorithm for Outlier Detection in Data Streams," International Journal of Advanced Research in Computer and Communication Engineering , pp. 3657-3666, 2013.

[18] B. Micenková, "Outlier Detection and Explanation For Domain Experts," Denmark, 2015.

[19] S. Guha, R. Rastogi and K. Shim, "ROCK : A Robust Clustering Algorithm for Categorical Attributes," Information System, pp. 345-366, 2000.

[20] D. L. Sreenivasa, M. N. Murthy and G. Athithan, "Outlier Analysis Of Categorical Data Using Navf," in IEEE Conference, 2013.

[21] R. P. Jakkulwar and R. Fadnavis, "Analysis of Outleir Detection in Categorical Data Set," IJERGS, pp. 622-625, 2015.

[22] Zhou, L. Wei, W. Qian and W. Jin, "HOT: Hypergraph-based Outlier Test for Categorical Data".

[23] D. Agarwal, "Detecting Anomalies in Cross-Classified Streams: a Bayesian Approach," Knowledge and Information System, pp. 29-44, 2006.

[24] S. S. Ahmed and H. Kitagawa, "Distance Based Outlier Detection on UNcertain Data of Gaussian Distribution," Springer, pp. 109-121, 2012.

[25] D. Ren, I. Rahal and W. Perrizo, "A Vertical Outlier Detection Algorithm with Clusters as by Product," in ICTAI, 2004.

[26] S. Ramaswamy, R. Rastogi and K. Shim, "Efficient algorithms for mining outliers from large data sets," in ACM SIGMOD, 2000.

[27] McCallum, K. Nigam and L. H. Ungar, "Efficient Clustering of High -Dimensional Data sets with Application to Reference Matching".

[28] Ghoting, S. Parthasarathy and M. E. Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets," in SIAM, 2006.

[29] T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in ACM SIGMOD, Canada, 1996.

[30] J. Zhao, C.-T. Lu and Y. Kou, "Detecting Region Outliers in Meteorological Data," in ACM GIS, New Orleans, Louisiana, USA, 2003.

[31] L. Akoglu, H. Tong, J. Vreekan and C. Faloutsos, "Fast and Relaible Anomaly Detection in Categorical Data," in ACM CIKM, Maui Hi USA, 2012.

[32] N. M.Hewai and M. K.Saad, "Class Outliers Mining: Distance - Based Approach," International Journal of Electrical and Computer Engineering, pp. 448-461, 2007.

[33] M. F. Jiang, S. S. Tseng and C. M. Su, "Two-phase clustering process for outliers detection," Pattern Recognition Letters, Elsevier, pp. 691-700, 2001.

[34] P. Filzmoser, R. Maronna and M. Werner, "Outlier Identification in high dimensions," 2006.

[35] K. Ro, C. Zou, Z. Wang and G. Yin, "Outlier Detection for High Dimensional Data," Biometrika, pp. 589-599, 2015.

[36] H. P. Kriegel, M. Schubert and A. Zimek, "Angle-Based Outlier Detection in High Dimensional Data," in International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008.

[37] Banarjee, "Density Based evolutionary outlier detection," in ACM GECCO, 2012.

[38] K. S., P.Visu and J.Janet, "A review on clustering and outlier analysis techniques in data mining," American journal of applied sciences, pp. 254-258, 2012.

[39] J. Laurikkalaa, M. Juholaa and E. Kentalab, "Informal identification of outliers in medical data," in IDAMAP, 2000.

[40] B. Gal, "Outlier detection," in Data Mining and Data Discovery Handbook, US, Springer, 2010, pp. 117-132.

[41] [Online]. Available: https://fhss.byu.edu/spss%20modeler/chapter%205.pdf.

[42] P. L. Rosin and F. Fiereins, "Improving Neural Network Generelization".

[43] P. Rana, D. Pahuja and R. Gautam, "A Critical Review on Outlier Detection Techniques," International Journal of Science and Research (IJSR) , pp. 2394-2404, 2014.

[44] J. Zhang, "Advancements of Outlier Detection: A Survey," ICST Transactions on Scalable Information Systems, pp. 1-26, 2013.

[45] N. Reunanen, "Modular framework for outlier detection," Finland, 2014.

[46] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques," IJCSI International Journal of Computer Science, pp. 307-324, 2012.

[47] M. VERLEYSEN, "Learning High Dimensional Data," Limitations and Future Trends in Neural Computation, pp. 141-162, 2003.

[48] V.Ilango, R. Subramanian and V. Vasudevan, "A five step procedure for outlier analysis in data mining," European Journal of Scientific Research, pp. 327-339, 2012.

[49] N. M.Hewai and M. K. Saad, "Class outliers mining: Distance-Based Approach," International Journal of Electrical and Computer Engineering , pp. 448-461, 2007.

[50] Zimek, R. Campello and J. Sander, "Ensembles for Unsupervised Outlier Detection: Challenges and Research Questions," SIGKDD Explorations, pp. 11-23.

[51] J. G. Cheng, "Outlier Management in Intelligent Data Analysis," London, 2000.

[52] M. O. Mansur and M. N. M. Sap, "Outlier Detection Technique in Data Mining: A Research Perspective," in Proceedings of the Postgraduate Annual Research Seminar 2005 , 2005.

[53] L. Sunitha, M. B. Raju and B. S. Srinivas, "A Comparative Study between Noisy Data and Outlier Data in Data Mining," International Journal of Current Engineering and Technology , pp. 575-577, 2013.

[54] V. Chandola, A. Banarjee and V. Kumar, "Outlier Detection: A Survey," 2004.

[55] S. Kim and S. Cho, "Prototype based outlier detection," in IJCNN, 2006.

[56] V. Schultze and J. Pawlitschko, "The Identification of Outliers in Exponenetial Samples," Statistica Neerlandica, pp. 41-57, 2002.

[57] E. Eskin, A. Arnold, M. Prerau, L. Portnoy and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data," Applications of Data Mining in Computer Security, 2002.

[58] L. Davies and U. Gather, "The Identification of Multiple Outliers," Journal of the American Statistical Association , pp. 782-792, 1993.

[59] Z. He, X. Xu and S. Deng, "An Otimization Model for Outlier Detection in Categorical Data," in Proceedings of ICIC, 2005.

[60] T. Cheng and Z. Li, "A Multiscale Approach for Spatio-Temporal Outlier Detection," Transactions in GIS, pp. 253-263, 2004.

[61] S. Shekar, C.-T. Lu and P. Zhang, "Detecting Graph Based Outlier: Algorithms and Applications," in ACM SIGKDD, 2001.

[62] S. Muthukrishnan, R. Shah and J. Vitter, "Mining Deviants in Time Series Data Streams," in SSDBM, 2004.

[63] Schlkopf, J. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the Support of a High Dimensional Distribution," Neural Computation, pp. 1443-1471, 2001.

[64] T. Hu and S. Y. Sung, "Detecting Pattern Based Outliers," Pattern Recognition Letters, pp. 3059-3068, 2003.

[65] J. Laurikkala, M. Juhola and E. Kentala, "Informal Identification of Outliers in Medical Data," in IDAMP, 2000.

[66] Yu, G. Sheikholeslami and A. Zhang, "Finding Outliers in Very Large Datasets," Journal of Knowledge and Information Systems, pp. 387-412, 2002.

[67] P. Rousseeuw and A.M.Leory, Robust regression and outlier detection, John Wiley and sons, 1996.

[68] H.S.Behera, A. Ghosh and S. K. Mishra, "A new hybrid K-means clustering based outlier detection technique for effective data mining," International journal of advanced research in computer science and software engineering, pp. 287-292, 2012.

[69] F. E. Grubbs, "Procedures for detecting outlying observations in samples," Technometrics, pp. 1-21, 1969.

[70] Y. Li and H. Kitagawa, "DB-Outlier Detection by example in high dimensional data sets," IEEE, pp. 73-79, 2007.

[71] N. Pham and R. Pagh, "A Near linear Time Approximation Algorithm For angle based outlier detection in high dimensional data," in KDD, 2012.

[72] Y. Li, D. Wu, J. Ren and C. Hu, "An improved Outlier Detection Method in High Dimensional Based on Weighted Hypergraph," in Second INternational Symposium on Electronic Commerce and Security, 2009.

[73] C. Aggarwal and P. S.Yu, "Outlier Detection for High Dimensional Data," in Proceedings of the ACM International Conference on Management of data SIGMOID, Santa BArbara, CA, 2001.

[74] S.D.Pachgade and S.S.Dhande, "Outlier Detection over data set using cluster based and distance based approach," International journal of Advanced Research in computer science and software engineering, pp. 12-16, 2012.

[75] P. Guo, J.-Y. Dai and Y.-X. Wang, "Outlier Detection in HIgh Dimension Based on Projections," in IEEE, Fifth INternational Conference on MAchine learning and cybernetics, 2006.

[76] R. Pamula, J. K. Deka and S. Nandi, "An outlier detection method based on clustrering," in IEEE, Second International conference on Emerging Applications of information technology, 2011.

[77] V. Barnett and T. Lewis, Outliers in statisticsl Data, Willy, 1994.

[78] D.M.Hawkins, Identification of outliers, Springer, 1980.

[79] R. Butler, "Outlier Discordancy Test in Normal Linear Model," JSTOR, Journal of Royal Statistical Society, pp. 120-132, 1983.

[80] S. S.S., "A Survey on Outlier Detection Methods," (IJCSIT) International Journal of Computer Science and Information Technologies, pp. 8153-8156, 2014.

[81] M. K. Deshmukh and A. S. Kapse, "A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach," International Journal Of Engineering And Computer Science , pp. 15453-15456 , 2016.

[82] M. Aouf and L. A. Park, "Approximate Document Outlier Detection Using Random Spectral Projection".

[83] Y. Zhang, N. Meratine and P. Havinga, "A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets".

[84] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," 2004.

[85] R. Kimball and M. Ross, The Data WareHouse Toolkit The Definitve Guide to Dimensional Modelling, John Wiley & Sons, Inc, 2008.

[86] M. H. Dunham, Data Mining Introduction and advanced topics, Printice Hall, 2002.

[87] M. Golfarelli and S. Rizzi, Data Warehouse Design Modern Principles and Methodologies, Tata McGraw Hills, 2012.

[88] P. Ponniah, Data warehousing Fundamentals A comprehensive Guide fro IT Professionals, 2006.

[89] W. Inmonn, 1996.

[90] Berson and S. J. Smith, Data Warehousing, Data Mining, and OLAP, Tata McGraw Hill, 2008.

[91] T. B. Pedersen, J. Thorhauge and S. E. Jespersen, "Combining.Data. Warehousing.and. Data.Mining.Techniques. for.Web.Log.Analysis," in Research and Trends in Data Mining Technologies and Applications, Monash University, Australia, Idea Group Publication, 2007, pp. 1-28.

[92] Jensen and J. Neville, "Data Mining in Social Networks," in Symposium on Dynamic Social Network Modeling and Analysis, 2002.

[93] P. Peng, Q. Ma and C. Li, "The Research and Implementation of Data Mining Component Library System".

[94] P. C. Agarwal, Probabality and statistics, 2007.

[95] Biswas, Probability and Statistics, 2012.

[96] J. L. Devore, Probability and Statistics for Engineers, 2011.

[97] A.Abede, J.Danials, W.McKean and J.A.Kapenga, Statistics and data analysis, Western Michigan University, 2001.

[98] S. Haykin, Neural Networks and Learning Machines, Pearson, 2008.

[99] V. Pudi and P. R. Krishna, Data Mining, Oxford University Press, 2009.

[100] R. Jindal and M. D. Borah, "A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS," International Journal of Database Management Systems (IJDMS) , pp. 53-74, 2013.

[101] Eskin, "Anomaly Detectopn over Noisy Data Using Learned Probability," in Machine Learning, 2000.

[102] W. Eberlea and L. Holder, "Anomaly detection in data represented as graphs," Intelligent Data Analysis, pp. 663-689, 2007.

[103] R. E. Marmelstein, "Application of Genetic Algorithms to Data Mining," in MAICS, 1997.

[104] R. O. Duda, P. E. Hart and D. G. Strork, Pattern Classification, Wiley, 2000.

[105] J. Singh and S. Agarwal, "Survey on Outlier Detection on Data Mining," International Journal of Computer Applications, pp. 29-33, 2013.

[106] M. K. Jiawei Han, Data Mining Concepts and Techniques, San Francisco: Morgan Kaufman , 2001.

[107] Smita and P. Sharma, "Use of Data Mining in Various Field: A Survey Paper," IOSR Journal of Computer Engineering (IOSR-JCE) , pp. 18-21, 2014.

[108] J. Xi, "Outlier Detection Algorithms in data mining," IEEE, Second International Symposium on Intelligent Information Technology Application, pp. 94-97, 2008.

[109] N. Padhy, D. P. Mishra and R. Panigrahi, "The Survey of Data Mining Applications And Feature Scope," International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), pp. 43-58, 2012.