CHAPTER 1
INTRODUCTION

1.1. OVERVIEW OF DATA MINING

A tremendous increase in the number of databases in each and every domain of human efforts has generated a larger demand for novel; potential tools for converting data into resourceful, task-based knowledge described Chen et al., (1996). In the endeavour to meet this requirement, researchers have been inquiring about concepts and techniques designed in machine learning, pattern identification, statistical data analysis, data visualization, neural nets, etc. These attempts have resulted in the evolution of a new research field, often referred to as data mining and knowledge discovery.

Data mining is described as the procedure of finding intriguing knowledge, like patterns, associations, modifications, anomalies and important structures, from enormous data that is stored in databases, data warehouses, or other information banks described Shyu et al., (2008). When data mining and Knowledge Discovery in Databases (KDD) are often considered to be synonyms, data mining is really a part belonging to the process of knowledge discovery. The figure (Figure 1.1) below illustrates the data mining in the form of a step in a repetitive knowledge discovery process.
Figure 1.1. Data Mining forms the basis of Knowledge Discovery process

The KDD process consists of a few number of steps beginning from collections involving raw data to some kind of new knowledge described Chen et al., (1998). The iterative process comprises of the steps mentioned below:

Data cleaning: also referred to as data cleansing, it is a stage where the noisy data and unnecessary data are excluded from the collection.

Data integration: at this phase, different data sources, mostly heterogeneous, may be integrated in a common source.

Data selection: in this step, the data having relevance to the analysis is determined and acquired from the data collection.

Data transformation: also referred to as data consolidation, it is a stage where the data selected is modified into forms suitable for the mining process.

Data mining: it is the critical step where intelligent strategies are used for extracting the patterns that have potential usability.

Pattern evaluation: in this step, highly fascinating patterns that represent knowledge are found on the basis of the measures given.
Knowledge representation: is the final stage where the identified knowledge is visually shown to the customer. This vital step employs visualization strategies to assist the users in understanding and interpreting the data mining results.

It is not uncommon to integrate few of these steps together. For example, cleaning and integration of data can be carried out at the same time in the form of a pre-processing stage in order to create a data warehouse. Selection of data and data transformation can also be integrated in which the data consolidation becomes the outcome of the selection, or, as in the case of data warehouses, the selection is carried out on the data transformed. The KDD is actually an iterative process. When the knowledge discovered is shown to the user, the assessment measures could be improved, the mining can be refined further, new data can be chosen or further modified, or new data sources can be combined, so as to make diverse, more suitable results to be available.

Data mining obtains its terminology from the similarities observed between looking out for useful information in a huge database and mining the rocks for getting a vein of precious ore. Both indicate either picking out through a huge quantity of material or intelligently probing the material in order to precisely spot the location of the values. However, it is actually misleading, as mining for gold in rocks is generally referred to as “gold mining” and not “rock mining”, therefore by similarity, data mining must have been rather referred to as “knowledge mining”. However, data mining went on to become the approved customary notation, and a very quick trend, which even eclipsed more generalized terms like knowledge discovery in databases (KDD), which specify a more absolute process. Other analogous terms that refer to data mining include: data dredging, knowledge extraction and pattern discovery.

1.1.1 DATA MINING MODELS

The data mining models consists of two variations: Predictive and Descriptive. The predictive model uses the known values to predict about unknown data values. Ex. Classification, Regression, Time series analysis, Prediction etc. The descriptive model discovers the patterns or associations in data and investigates the
characteristics of the data analysed. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

Several of the data mining applications are targeted at predicting the future state of the data. Prediction is the procedure of evaluating the present and past states of the attribute along with the prediction about its future state. Classification is a method representing the mapping of the target data onto predetermined groups or classes, this is actually a supervised learning as the classes are predetermined in prior to the analysis of the target data. The regression is involved with the learning of function, which maps the data item onto absolute valued prediction variable. In the time series analysis, the value of an attribute is analysed as it changes with the passage of time. In time series analysis, the distance measures are utilized to decide about the similarity between various time series, the structure of the line is analysed in order to decide about its behavior and then the historical time series plot is employed for predicting the future values of the variable.

Clustering is identical to classification excepting that the groups are not predetermined, but are specified by the data only. It is also known as unsupervised learning or segmentation. It refers to the division or classification of the data into groups or clusters. The clusters are specified by examining the behavior of the data carried out by the domain specialists. The term segmentation is employed in a very particular context; it is actually a procedure of segmenting the database into disjoint grouping consisting of identical tuples. Summarization is defined as the method of exhibiting the summarized information obtained from the data. The association rule discovers the association between the various attributes.

Association rule mining is actually a two-step process: Discovering all the frequent item sets, creating potential association rules from the frequent item sets. Sequence discovery is defined as the process of discovering the sequence patterns present in data. This sequence can be utilized for understanding the trend.
1.2. APPLICATIONS OF DATA MINING

Data mining methodologies are implemented in various decision-making scenarios in organizations. Its significant cannot be underscored, since it applies in various fields few of which are discussed as follows.

Marketing: This consists of the evaluation of customer behavior in purchasing trends, market strategies determination ranging from advertising to location, targeted mailing, customer segmentation, products, stores, catalogue design and advertising strategy.

Supply chain visibility: Companies have made parts of their supply chain to automatic, facilitating the gathering of important data regarding inventory, supply performance and logistic of materials, and finished goods, material costs, accuracy of plans made for order delivery. Data mining application also cuts through cost optimization and work force evaluation in organizations.

Geospatial decision making: In the scenarios involving climate data and earth ecosystem, automatic extraction and evaluation of intriguing patterns that involve themodelling of ecological data and developing effective algorithm for getting spatio temporal patterns in the form of tele-connection patterns or iterative and consistent climatic patterns. This operation is generally performed employing the clustering method that segments the data into useful groups, aiding in the automation of the revelation of tele-connections.

Biomedicine and science application: Biology remained a field ruled by an attitude of formulating hypothesis, conducting experiment, evaluating the results, but now due to the influence of data mining it has developed into a field of big science attitude that involves the collection and storing of data, mining new hypothesis, then assuring with data or additional experiment. It also consists of discovering patterns in radiological images, analysing microarray (gene-chip) experimental data, so as to cluster the genes and to create association with symptoms or disease, evaluating the side effects of drugs and efficiency of particular drugs.
Manufacturing: In this aspect, the application corresponds to the optimization of the resources utilized in the optimal design of manufacturing processes, and product design on the basis of customers feedback.

Telecommunications and control: It is applied over the extensively available voluminous data comprising of call records and other telecommunication relevant data that, in turn, is used in toll-fraud discovery, consumer marketing and improvement of services.

Data mining is also used in security operations and services, information analysis and delivery, text and Web mining cases, banking and commercial applications and also insurance.

1.3. OVERVIEW OF CLUSTERING

Being one the most essential tasks in data mining, clustering targets at grouping a set of objects into clusters so that the objects present within the same cluster have more similarity with one another compared to objects present in other clusters described Feng et al., (2016). When the clusters are determined, then the objects are marked with their respective clusters, and the general characteristics of the objects in cluster are summarized in order to create a class description. The points discussed below, focus on the need of clustering in data mining.

Scalability: requires greatly scalable clustering algorithms to tackle with huge databases.

Capability to handle various types of attributes: Algorithms must have the ability to be used on any variety of data like interval-based (numerical) data, categorical, and binary data.

Discovery of clusters with attribute shape: The clustering algorithm must have the capability of finding clusters of random shape. They must not be restricted to just distance measures, which tend to discover only spherical cluster of tinier sizes.
High dimensionality: The clustering algorithm must have the capability to deal with low-dimensional data and also the high dimensional space described Fred and Jain (2005).

Ability to deal with noisy data: Databases may have noisy, missing or incorrect data. Few algorithms have sensitivity to such kind of data and may result in poor quality clusters.

Interpretability: The clustering results must be understandable, meaningful, and resourceful.

1.3.1. Clustering methods

The current clustering algorithms can be generally divided into five types: Partitioning technique, Hierarchical technique, Density-based technique, Grid-based technique, Model-based technique which are discussed below.

**Partitioning technique:** Generate a segment of a dataset D consisting of n objects in a set with k clusters. Partition based algorithm are k-means and k-medoids described Huang et al., (2005). In k-means, every cluster are indicated by means of the center of the cluster. The variant of k-means algorithm is k-modes, which cluster categorical data by substituting the mean of cluster with modes. K-medoids algorithms are PAM, CLARA, and CLARANS in which every cluster is denoted by one among the objects present in the cluster. Vijayarani and Nithya (2011) introduced a partitioning clustering algorithm for detecting outliers.

**Hierarchical technique:** Create a hierarchical division of the set of data using some criteria. Hierarchical technique is divided into agglomerative or divisive depending on whether hierarchy is formed in top-down or bottom-up form. Agglomerative is bottom-up strategy that merge cluster into larger cluster until all object are into one cluster or until some condition for termination are satisfied. Divisive is top-down strategy and is reverse of agglomerative which subdivide the cluster into small cluster. For given input set S, the goal is to produce hierarchies in which nodes represent subset of S. This method form tree structure of cluster. Every level of the tree denotes a partition of input data into several cluster or group.
Hierarchical clustering algorithm is Balance Iterative Reducing and Clustering using Hierarchies (BIRCH), Cluster Using Representatives (CURE). Strength of this method is no need to assume or define number of cluster initially.

**Density-based method:** Many of the partition based clustering methods are dependent on distance. Distance based method deal with only spherical-shaped cluster and difficult for arbitrary shapes. Density based method use connectivity and density function to make cluster. Density-based algorithm includes Density-Based Spatial Clustering of Application with Noise (DBSCAN). This algorithm is used to discover cluster of arbitrary shapes.

**Grid-Based Method:** Assign object to the suitable grid cell and calculate the density of every cell. Remove cells, whose density falls below a specific threshold. Advantages of this method are fast processing, independent of the number of data object and there is no need of distance computation. This method is depends on number of cells. Grid-based algorithms are Statistical Information Grid approach (STING) and Clustering in Quest (CLIQUE).

**Model-Based Method:** Model-based techniques provide the hypothesis of a model for every one of the cluster and gets the best fit of the data to the model given. This technique discovers the clusters through the clustering of the density function. It shows the spatial distribution of the data points. This technique also offers a means for automatically determining the number of clusters on the basis of standard statistics, considering outlier or noise. Therefore, it makes a reliable clustering technique. Model-based technique is Expectation-Maximization (EM).

### 1.4. OUTLIER DETECTION

An outlier is basically an observation point, which is located at a distance from the other observations described in Srivastava (2005). An outlier is seen because of a variation in the measurement or it might represent experimental error and sometimes the latter gets eliminated from the data set.
Being the observations at the most extreme, outliers may consist of the sample maximum or sample minimum or both based on whether they are highly high or low. The sample maximum and minimum are not outliers at all times as they may not be generally distant from the other observations.

Types of outliers

Generally, outliers can be divided into three groups, referred to as global outliers, contextual outliers and collective outliers, as shown in figure 1.2.

![Diagram of Outliers](image)

**Figure 1.2. Types of Outliers**

1) Global outlier

In a given dataset, a data object can be considered to be a global outlier when it considerably deviates from the remaining data set. Sometimes, the global outliers are known as point anomalies and fall under the simplest kind of outliers. Many of the outlier detection techniques are targeted at getting global outlier.

2) Contextual Outliers

In a given dataset, a data object is referred as a contextual outlier when it has significant deviation with regard to a particular context of the object. Contextual outliers are also referred to as conditional outliers since they are conditional on the chosen context.

3) Collective Outliers
A subset of data objects deviate together considerably from the entire data set, even when each of the data objects might not be outliers. Detecting the collective outliers, take not just the behavior of individual objects into consideration, but also the collection of objects. It is required to possess the background information regarding the association among data objects, like a distance or similarity measure on objects.

1.4.1. Outlier detection methods

The different outlier detection approaches shall be categorized into: Statistical Distribution-Based method, Distance-based method, Deviation-based method, density-based method.

![Outlier Detection Approaches](image)

**Figure 1.3. Outlier Detection Approaches**

**Statistical distribution-based Outlier detection**

This method presumes a distribution or probability model of a dataset given and identifies outlier by using discordancy test. Discordancy test depends on data distribution, distribution parameter and number of expected outlier. There are certain
kinds of statistical distribution such as Gaussian. In which parameters are calculated by assuming that every data point is formed by statistical distribution such as mean and standard deviation. Outlier is the point that has low probability to be formed by overall distribution. Disadvantage of this method are most tests are for single attribute and require the knowledge about data distribution parameter.

**Distance-based Outlier detection**

This approach introduces to overcome the primary disadvantage of previous statistical approach, which is this approach work with multi-dimensional analysis. A distance based outlier can be define as, An object o, in the dataset, D is actually a distance-based outlier with parameters pct and $d_{\text{min}}$, implying, a DB (pct, $d_{\text{min}}$)- outlier, if it is at least a fraction, pct, of the object in D remain at a distance more than $d_{\text{min}}$ from o. To find Distance between point with its neighbor, the different dissimilarity measure used are Euclidean distance, cosine distance, city block distance, etc. Pachgade and Dhande (2012) used k-means algorithm for the formation of cluster and distance-based outlier detection approach to detect outliers. They use Euclidean distance for dissimilarity measure.

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)$$

(1.1)

Where, $\bar{x}_r = \frac{1}{n} \sum x_{rj}$ and $\bar{x}_s = \frac{1}{n} \sum x_{sj}$

Pamula et al., (2011) introduced a k-means algorithm to cluster the dataset and use Local Distance-based Outlier Factor (LDOF) to detect outlier.

**Deviation-based Outlier Detection**

This approach identifies outlier by observing main characteristics of object in data set. The object that deviates too much from these feature are consider as outliers. Two technique employed for deviation-based outlier detection include sequential exception approach and OLAP data cube approach. Sequential exception technique selects the series of subsets from the set for the purpose of analysis and determines dissimilarity difference for each subset according to the previous subset in
the sequence. OLAP data cube method utilizes data cube for identifying the regions of anomalies in huge multidimensional data.

**Density-based Outlier Detection:**

Distance-based outlier detection approach exhibit problem with various densities. The basic idea of this technique is the comparison of the density around the point with the density computed around its adjacent neighbors. The relative density of a point in comparison with its neighbors is calculated to be an outlier score. The density found around a common data object is identical to the density found around its neighbors. The density found around an outlier is dissimilar to the density found around its neighbors. To define Local Outlier Factor (LOF), it need the concept of k-distance, k distance neighborhood, reach ability distance, and local reach ability density. The reach ability distance of an object p corresponding to object 0, is define as reach_dist_{minpts}(p,o)=max{MinPtsdistance(o),d(p,o)}. Local reach ability distance (lrd) of point p, inverse of the average reach-dists of the kNNs of p,

\[
Lrd(p) = \frac{|N_{minpt}(p)|}{\sum_{o\in N_{minpt}(p)} \text{reach-dist}(p,o)}
\]  

(1.2)

and Local Outlier Factor (LOF) of p, average ratio of neighbors of p and lrd of p

\[
LOF(p) = \frac{\sum_{o\in N_{minpt}(p)} \frac{lrd(o)}{lrd(p)}}{|N_{minpt}(p)|}
\]  

(1.3)

It can be determined if a point p is a local outlier depending on the computation done of \(LOF_{\text{MinPts}}(p)\). Wang and Su (2011) introduced a density-based outlier detection technique. They first apply density-based technique to remove noise data and then apply k means to cluster data.
1.4.2. MEASUREMENTS FOR OUTLIER DETECTION

Outlier detection is a research challenge with high importance in data mining, whose objective is the discovery of resourceful abnormal and uneven hidden patterns in huge datasets.

Entropy and Total Correlation

Suppose a set $X$ consisting of $n$ objects $\{x_1, x_2, \ldots, x_n\}$ each $x_i$ for $1 \leq i \leq n$ being a vector with categorical attributes $[y_1, y_2, \ldots, y_m]^T$, where $m$ refers to the number of attributes, $y_j$ has a value domain specified by $[y_{1,j}, y_{2,j}, \ldots, y_{n,j}]$, $j = 1, \ldots, m$ and $n_j$ refers to the number of unique values in attribute $y_j$. Assuming every $y_j$ to be a random variable, the random vector $[y_1, y_2, \ldots, y_m]^T$, is denoted by $Y$. $x_i$ can be represented as $\{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$. $H_x$, $I_x$ and $C_x$ are used, respectively, to refer to entropy, mutual information, and total correlation calculated on the set $X$; e.g., $I_x(y_i, y_j)$ indicates the mutual information between attributes $y_i$ and $y_j$ described Gaddam et al., (2007). At times, the index term $X$ is dropped off if no ambiguity exists, e.g., applying $(y_i; y_j)$ instead of $I_x(y_i, y_j)$. So, depending on the chain rule for entropy, the entropy of $Y$, represented as $H_x(Y)$ can be written as below:

$$H_x(Y) = H_x(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H_x(y_i | y_{i-1}, \ldots, y_1) \tag{1.4}$$

Where,

$$H_x(y_m | y_{m-1}, \ldots, y_1) = - \sum_{y_m | y_{m-1}, \ldots, y_1} p(y_m, y_{m-1}, \ldots, y_1) \log p(y_m, y_{m-1}, \ldots, y_1)$$

The entropy can be employed in the form of a global measure in outlier detection. With regard to information theory, entropy implies uncertainty corresponding to a random variable: when the value of an attribute is not known, the entropy of this attribute represents how much of information needs to do the prediction of the right value. A subset $n$ consisting of objects is good outlier candidates in case their elimination from the data set leads to a considerable reduction.
in the entropy of the data set. The technique introduced in Lee and Xiang (2001) uses entropy in the form of a quality measure in the detection of outlier from unidimensional audio data.

**Total Correlation**

He et al., (2005) extended this schema in order to carry out the measurement of the disorder of a multidimensional data set removing the outliers, in which a heuristic local search is used for reducing the objective function. The following details on how the total correlation can also be utilized in outlier detection. The total correlation is defined to be the summation of mutual information of multivariate discrete random vectors $Y$, represented as $C_X(Y)$

\[
C_X(Y) = \sum_{i=2}^{m} \sum_{(r_1,...,r_i) \in \{1,...,m\}} I_X(y_{r_1},...,y_{r_i})
\]  

(1.5)

where $r_1 \ldots r_i$ refer to attribute numbers selected from 1 to $m$. $I_X(y_{r_1};\ldots;y_{r_i}) = I_X(y_{r_1};\ldots;y_{r_i-1}) - I_X(y_{r_1};\ldots;y_{r_i-1}|y_{r_i})$ stands for the multivariate mutual information of $y_{r_1} \ldots y_i$, where $I_X(y_{r_1};\ldots;y_{r_i-1}|y_{r_i}) = E(I(y_{r_1};\ldots;y_{r_i-1})|y_{r_i})$ is the conditional mutual information. The total correlation is actually a quantity, which measures the mutual dependence or shared information belonging to a data set described He et al., (2005).

Consider the case of total correlation $C_X(y_1; y_2)$ having two attributes $y_1$ and $y_2$ in the form of an example, $C_X(y_1; y_2) = I_X(y_1; y_2)$ represents the total correlation for a random vector $Y$ having two attributes $y_1$ and $y_2$. Its value is associated with the decrease in the uncertainty of one attribute value rendered by the knowledge of another. In case the value of $C_X(y_1; y_2)$ is big, it indicates that the number of duplicate pairs of attribute values is smaller in these two attributes in comparison with the condition when the value of $C_X(y_1; y_2)$ is tiny. On the whole, for the scenario where there are attributes more than two, greater $C_X(Y)$ indicates a smaller number of objects that share common attribute values that, in turn indicates a lesser number of frequent item sets and a not good cluster structure. Therefore, like
entropy, the total correlation can be employed for measuring the goodness of the outlier candidates in a subset O by assessing \( C_X(Y) \) for \( X=X\backslash O \). Moreover, the lesser the value of \( C_X(Y) \), then the better would be the subset O in the form of a set of outlier candidates.

**Holoentropy**

The system begins with an example to prove that just entropy is not a sufficient measure for outlier detection and the contribution made by total correlation is also required. As observed from the example, in which 14 objects with four attributes are shown, the data set is represented by X. X consists two objects \( x_{13} \) and \( x_{14} \) that can be detected to be the most possible outliers by comparing with the remaining 12 objects. In addition, \( x_{14} \) is obviously more exceptional compared to \( x_{13} \) as it does not share any of its attributes with the remaining objects. Now, \( H_{X\backslash x_{14}}(Y)=H_{X\backslash x_{13}}(Y)=3:7 \) indicates that, if just the entropy is employed, \( x_{14} \) and \( x_{13} \) are exceptional on an equal level like outlier candidates.

On the other side, in case the total correlation and the entropy are combined, \( H_{X\backslash x_{14}}(Y)=H_{X\backslash x_{13}}(Y)=9:414 \) and \( H_{X\backslash x_{13}}(Y)=H_{X\backslash x_{13}}(Y)=9:414 \) is obtained that lets object \( x_{14} \) to be differentiated as a more possible outlier compared to \( x_{13} \). It is rather interesting that, with the given distributions of attributes in a data set, there exists a complementary association between the entropy and total correlation of Y.

Definition 1: (Holoentropy of a random vector). The holoentropy \( HL_X(Y) \) is defined to be the summation of the entropy and the total correlation of the random vector \( Y \), and can be defined by the summation of the entropies on every attribute,

\[
HL_X(Y) = H_X(Y) + C_X(Y)
\]  

(1.6)

**Drawbacks in Outlier Detection**

 Speaking in an abstract manner, outliers are generally patterns deviating from anticipated normal behaviour that can be denoted in its simplest form by a region and all the normal observations are visualized to be in this normal region and the remaining is considered as outliers. This technique seems simple but is a huge
challenge owing to reasons below. It is very hard to specify either the normal behavior or a normal region. The hardships faced are listed below.

Inclusion of each of the likely normal behavior in the region.

Inaccurate boundary between normal and outlier behaviour because sometimes outlier observation existing near to the boundary could really be normal, and vice-versa.

Adaptation of dangerous intruders to show the outlier observations to seem normal while the outliers are due to suspicious actions.

In several domains, the term of normal behaviour keeps on changing and might not be updated currently in order to be one representative in the future.

Varying conceptions about outliers in diverse application domains renders it hard to use the method evolved in one domain to the other one. For instance, in the medical domain, a deviation seen in a small amount from the usual body temperature may be an outlier, whereas similar kind of value deviation in the stock market domain may be regarded to be normal. Even inside the same domain, for example, crime detection, scenarios may be such that where usage of foreign made weapons might be regarded normal in crimes that are committed in metro cities but will be treated an outlier for the commoner murders in tribal areas.

1.5. PROBLEM SPECIFICATION OF OUTLIER DETECTION METHODS

The recent developments in the area of data mining have resulted in the outlier detection process. Outlier detection targets at finding patterns in data, which do not adhere to predictable behavior. The number of outlier detection techniques has been proposed that use different mechanism such as Local Distance-Based Outlier Factor, fuzzy clustering, k-mean clustering and neural networks. These are tried to reduce the impact of outliers or remove. The datasets can be constant having a small number of attributes in which outlier detection is considerably easy. However, the datasets can also be dynamic, like data streams, and simultaneously possess a big number of attributes. Handling this type of datasets is more complicated by nature and
needs paying more attention to the performance of detection of the techniques to be designed.

1.6. OBJECTIVE OF THE WORK

The important goals of this work are

- To find the outlier attribute accurately
- To overcome the Imbalance data problem
- To minimize the time and computational complexity
- To enhance the outlier detection performance using BAT optimization algorithm
- To design a system for dealing with the missing values

1.7. RESEARCH CONTRIBUTION

The prime contribution of this work is focused on outlier detection. This research scope extends to

1. Ascent-based montecarlo expectation– maximization outlier detection used for large-scale categorical data
2. Single pass attribute value frequency based outlier detection employed for large-scale categorical data
3. Efficient outlier detection using graph based semi supervised clustering with bat algorithm

1.8. OVERVIEW OF THE THESIS

This chapter 1 studies about the overview over data mining along with its functions and applications. The various types of outlier and outlier detection methods are explained in detail. The outlier detection is used in various fields such as marketing, supply chain visibility geospatial decision making, and biomedicine and science application. For the purpose of improve the outlier attributes detection performance better mechanism is needed.
Chapter 1 discusses about the introduction of data mining and its applications, clustering, outliers, types of outliers and outlier detection approaches. The main objective of this research is also been discussed briefly.

Chapter 2 overviews the short discussion of the relevant works on various measurement functions, clustering, and optimization algorithm based outlier detection.

Chapter 3 explains novel approach which combines the attributes based Kullback-Leibler divergence (KLD) for attribute weighting process and performs the Ascent-based Monte Carlo expectation–Maximization (AMCEM) methods for outlier detection in detail. This chapter discusses in detail about the overall proposed methodology and is compared with the existing approach.

Chapter 4 explains a SinglePass Attribute Value Frequency (SAVF) which is used for outlier detection. It is also discussed about the overall proposed methodology and comparison of existing as well as proposed methodologies significantly.

Chapter 5 discuss about an efficient outlier detection using graph based semi supervised clustering with bat algorithm. This chapter carries out the comparison of the overall performance between the newly introduced techniques.

Chapter 6 discuss about the experimental results and performance evaluation. This chapter compares the overall performance of the proposed approaches.

Chapter 7 provides the conclusion which explains about graph based semi supervised clustering with bat algorithm achieves better detection rate. In addition, the future scope for improvement is also dealt with in this chapter.