International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2020) pp. 398-411

Received: 27 May, 2019; Accepted: Accepted 19 Oct, 2020; Publish: 3 December, 2020

A Study on Different Methods of Outlier Detection Algorithms in Data Mining

T. Sangeetha [1], and Geetha Mary A [2]

[1] School of Computer Science and Engineering, Vellore Institute of Technology,
Vellore - 632 014, Tamilnadu, India
sangee_arasu05@yahoo.co.in

[2] School of Computer Science and Engineering, Vellore Institute of Technology,
Vellore - 632 014, Tamilnadu, India
geethamary.a@gmail.com

***Abstract*: In the modern world, data are available widely, and it is essential to transform such data into useful knowledge and information. Data mining has attracted considerable awareness in the field of information and its community. Dataset is a collection of significant objects which do not belong to the same category. Some objects differ slightly from other regular objects are identified as outliers. Detecting outliers are notable because its presence slows down the system performance. Most methods of data mining dismiss outliers as noise or exceptions. However, rare events can be more attractive in some applications, such as fraud detection than frequent events. Outlier analysis is also known as outlier mining. Many fields such as marketing, sales, production, fraudulent identification, customer retention, and scientific research, use the acquired data. Rough sets are used to handle uncertain and vague data present in the real world. The discussion of rough classification, clustering, and different outlier detection methods are carried out in detail with suitable algorithms and examples. This survey provides an overview of outliers and existing outliers by classifying them into different dimensions. Wine dataset from the UCI repository has been taken to prove the performance of the rough set based entropy measure with weighted density value over existing methods.**

Keywords: data mining, fraud detection, outliers, outlier detection methods, outlier mining, rough sets**.**

I. Introduction

Data can be any subject, number, or content that is quickly processed by a system. Today, companies have an enormous amount of data in different styles and aspects. It includes operational information such as inventory and financing, non-operational information such as weather forecasting, financial information, and meta-information, such as the design of different databases or definitions for a word like a dictionary [1]. Data modeling or link between these objects supplies some information. Based on the facts of the past, obtain knowledge from the information. The point-of-sale system has improved by customers purchasing behavior. In recent years, the accumulation of data is more by the mass data acquisition in supermarkets, images produced by satellites, and data in networking systems.

It is like drinking from a fire hose to get information out of the data. The sheer data size outweighs our capability to search for data manually in the hope of finding useful information. So, researchers have adopted data mining techniques to do fast analysis and summarization of data. In brief, the main task of data mining is to discover elegant knowledge automatically from repositories, which are substantial [2]. The perspective of every person is unique. Each data provides different meanings for different users. In the view of a business perspective, managers or analysts use it to make profitable companies. Earth scientists use it to retrieve knowledge about unknown data. System administrators use it to detect unauthorized users from resource access. The elemental part of Knowledge Discovery Databases (KDD) is data mining. It includes three steps, such as preprocessing, mining of data, and post-processing. The preprocessing stage involves the correcting of data in the right format and selecting relevant attributes to make them ready for further analysis [3].

After processing, all operations are carried out to provide the results of data mining, which are easy to accept and interpret. The output can be ordered or filtered in such a way to eliminate uninteresting patterns to eliminate different measures. Techniques like visualization are used to provide scientists for investigating the results obtained in data mining. The classification of data mining functionalities is predictive and descriptive. In predictive, some variables are used to predict the variables of another. E-tailers predict their online customers who buy their product, biologists predict protein function task, and analyst of the stock market predicts the prices of future stocks. In a descriptive task, interpretation of patterns has been made manually, and its relationships have been discussed in detail.

For instance, Earth scientists need to interpret which force influencing climate patterns. In intrusion detection, programmers have to identify who attacks the systems in networks. Then while analyzing documents that group shares, the common theme needs to be interpreted. The predictive model is used to extract data. The type of input given to predictive analysis is variables that are explanatory to define the data and target values with prediction. In online shopping, critical customer data such as age, salary, gender, location, and how they are accessing a web page for how much time [4]. Except for this, one more variable buy is used to predict whether the customer buys or not.

Predictive modeling can further extend to two kinds, like classification and regression. Clustering is a technique to find similar data items which are grouped together. Data items in one cluster may be similar to data items in another cluster. It is used to segment customers on market, categorize documents or segment land according to its vegetation limit. It gives better understanding and description of data also useful when data sets are large. Some objects belong to single cluster only. They are replaced as representative object. The reduced set of representative objects can be used further in analysis of data. When objects characteristics or behavior significantly deviates from other objects, they are known as anomaly detection. It has been used in intruder system, and for predicting fraudulent activities of credit card system. Other approaches for anomaly detection are based on statistics or inter space or graphical concepts.

A.Classification

Based on the given input, classification predicts the output. The algorithm needs attributes for training and to predict the target attribute. The goodness of this algorithm has been known and how it analyzes the input and the outcome is forecasted [5]. The input of the training set (Table 1) is the patient dataset, which was recorded earlier, whether the patient has the symptoms of heart problem or not, and the prediction set (Table 2) has clearly shown.
Classification techniques use prediction rules to acquire knowledge. Prediction rules are generated in the form of IF-THEN, whereas the IF part defines conditions based on requirement, and the THEN part produces the prediction attribute.

Age	Pulse	BP	Heart Problem
45	99	150/90	?
67	56	110/65	?
85	75	153/75	?

Table 1: Training Set.

For instance, IF (Age=65 AND Pulse>70) OR (Age>60 AND BP>140/70) THEN there is a chance to get heart problem, which represents "yes." The classification method uses conjunctions like AND, OR, which provides a relationship of attributes so that it is easy for an analyst to predict.

B. Clustering

Classification analyzes class labeled data to train and predict the future, whereas clustering analyzes without considering class labels [6]. Clusters are a grouping of objects which are having a high degree of similarity or similar objects within a class or intraclass. Likewise, objects in a dataset which are having similar behavior or characteristics form different groups or clusters. Consider an electronic shop's customer data to distinguish different subpopulations of customers. Figure 1 shows the identified target groups of marketing.

C. Outliers or Anomalies

Objects which behave differently from our expectations are anomalies or outliers. Detection of outliers is vital in the field of medicine, damage of equipment in industries, people's safety and surety, video surveillance and also to detect intruders malfunction [7]. The concepts of outliers and clusters are the two which related profoundly. Clustering technique structures the data according to the majority patterns available in the dataset. However, outlier analysis tries to identify the exceptional cases which significantly deviates from the majorities. But clustering technique and outlier analysis offer different services.

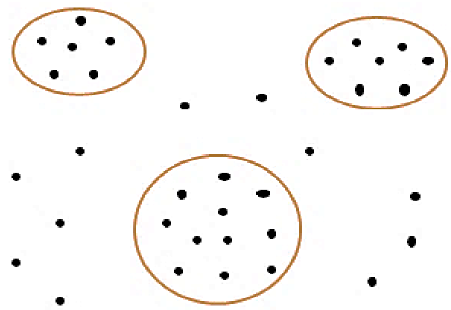

Figure 1. Data clusters in a specific city

D. Applications of Data Mining

The marketing field mainly uses data mining techniques. It needs past, current, and predictive data to know about its customers, sales, and competitors. Without data mining techniques, it is impossible to identify their strengths and weakness, feedback given by the customers, and also to make effective decisions during an emergency period. Some of the analytical processing tools are needed to store and process multidimensional data.

Classification plays a vital role in predicting its customers, sales, and product supply, whereas clustering provides a grouping of data under similar characteristics. A search engine is used to collect information on the web. Based on the user query, results are produced or how many users hit a particular web page, which is known as hits. Data can be in any form, such as text, images, web pages, or excel sheets. Also, data have retrieved through public repositories [8].
Search engines used on the web are computer servers specializing in gathering information about the web access and storage of data. User query search results frequently returned; some lists are known as hits. It may be multimedia files or even other file formats where searches have done manually or algorithmically. Figure 2 shows the different types of data mining techniques.

E. Structure

This article provides different techniques for outlier detection, which are structured as follows: A detailed description of outlier analysis has made in Section II. Section III clearly explained outlier types and their classification. Section IV discusses the different approaches for detecting outliers.

Section V and VI describe parametric and nonparametric methods for outlier detection. Section VII describes outlier detection based on proximity, and Section VIII describes outlier detection based on clustering. Rough sets classification, clustering, and outlier detection for numerical and categorical data described in Section IX and the final section end up with the conclusion.

II. Outlier Analysis

Consider an example, a bank which issues credit card monitors customer purchasing behavior frequently. It mainly focused on money on how much they are spending and their location. If something violates, like purchasing for a more significant amount than usual and location different from their residence, denotes misuse of the credit card by some other person.

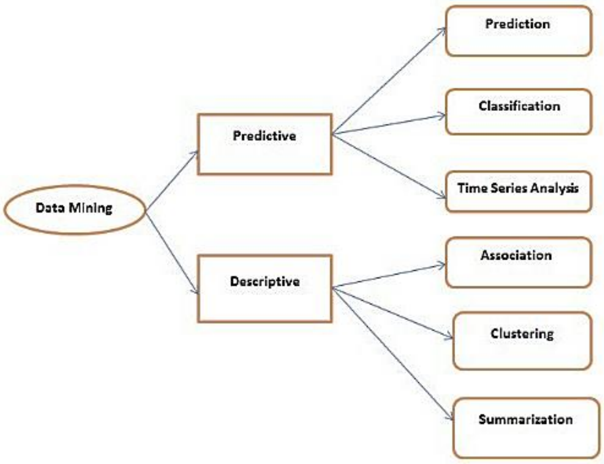

Figure 2. Variety of Data Mining Techniques

These types of transactions have been monitored because they differ from the usual ones. The patterns of transactions have been differed when the credit card has stolen. These kinds of objects which diverge from their normal behavior or pattern are outliers [9]. Suppose if a customer buys extra coffee rather than usual or buying some bulk items may not be denoted as outliers. And if a company sends false alarm, always customers get dissatisfied. The outlier differs from noise. Many data mining process involves removing noise before it has been processed. Suppose a new content appeared in social media, it may be defined as outliers initially, but later, when it goes under the general category, it acts as a regular data. Figure 3 shows outlier objects in region X.

A.Outlier Classification

The classification of outliers is of three different types, such as global, contextual, otherwise known as conditional or collective anomalies.

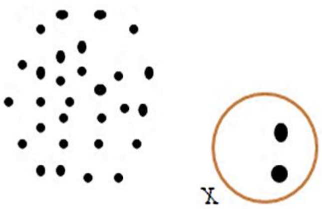

Figure 3. The objects in region X are outliers

1) Global Anomalies

Typical changes in patterns generated while evolving are known as global anomalies. Suppose in a trading system, transactions that do not obey a universal rule or if a hacker tries to hack some systems, broadcasting of messages may be within a short period are examples for global outliers [10].

2) Contextual or Conditional Anomalies

These kinds of anomalies exist only in a particular context. For example, the temperature recorded at 28°C in Chennai is abnormal during the winter season but appeared to be usual during the summer season [11]. For that day alone or a particular context, it differs is called contextual or conditional anomalies. The reason is that they are constrained to a particular context. It is of two attributes, such as contextual and behavioral attributes. In the example, the temperature is contextual based on place and date and behavioral concerning pressure and humidity. This method is the generalization of local outliers that detect outliers in density-based. It happens to be within the local region where it occurs. Global outliers consider a whole dataset as a context that holds empty contextual attributes.

3) Collective Anomalies

The group of objects deviates from the whole dataset, forms collective outliers. For example, the supply department handles thousands of orders and transactions per day. If anyone of the delayed shipment, it is general [12]. Because minimal delay is common and acceptable, but if more than 100 orders delayed at a time, it results in a drastic change that forms collective outliers. If a computer system generates a service denial message is not a problem. But the group of systems sends denial messages to each other indicate hacking, which results in collective outliers. Another instance, if a transaction occurred between two parties, is standard, but the bulk of the stock exchanged within a short delay is a notable one, which results in collective outliers. In general, global outliers are simple and easy to detect. Contextual outliers needed background data based on attributes such as contextual and behavioral. Collective outliers also needed background data to build a relationship between objects to form groups.

B. Methods of Outlier Detection

1) Supervised Technique

This method models normal and abnormal data. Data miner analyzes, and the label assigned to each sample of data. Then outliers are detected under the problem of classification. A classifier is used to train and test data [13]. The standard data are labeled, and which does not match the model are outliers. The challenges to supervised outlier detection techniques are: It uses two classes such as regular and outliers. Outliers are generally available in small numbers when compared to ordinary objects. So imbalancing technique is used to do oversampling. Artificial outliers are generated to make distribution equal to put into a classifier and mislabeling of natural objects as outliers should not be happened. Experts have to analyze it carefully before labeling each object.

2) Unsupervised Technique

In many of the applications, labeling of objects such as regular or outliers are not available. Instead, they are assigned implicitly. Clustering is a technique where objects fall under similar features forms a group[14]. A different object which is distant from similarity measures is outlier. The challenges faced here are, objects which do not fall any category may not be an outlier. Instead, they might be noise. High cost requires to construct clusters and then to detect outliers.

3) Semi-Supervised Technique

It is a combination of supervised and unsupervised techniques. Most of the applications need labeled data and unlabeled data [15]. A small number of objects are outliers, and the remaining are general objects. These labeled outlier objects act as a model for remaining outlier objects to detect.

C. Challenges in Outlier Detection

The modeling of normal and outlier objects should be sufficient. Building an extensive model for standard data is a challenging task, and it is complex to define all the possible behaviors which are normal. Not all the cases labeling of data as “normal” or “outliers” work. Sometimes the generated scores also used to identify outliers [16].

The outlier detection method varies according to an individual application. So constructing a universal method for outlier detection is impossible and application dependent. Some need similarity or distance measures, and others use a relationship model to identify outliers [17]. For example, a deviation in the clinical test report is natural to detect outliers, but in marketing, fluctuations of data are high. So, identifying outliers are complex.

Already we discussed that noise is different from outliers. Low-quality data and the presence of noise are the most challenging in outlier detection. They can malformed the data and also cause blurriness to differentiate between regular and outlier objects [18]. Some noise or missing data may be identified as outliers mistakenly, which reduces the effectiveness of outlier detection.

Researchers are interested not only in detecting outliers but also to find the reason why they are treated as outliers. For instance, if researchers used any statistical methods, how much degree it deviates from the other objects to be shown [19]. The same method applies to the whole dataset; objects which likely deviate from others are outliers.

D. The different outlook of the outlier detection problem

Specific issues can be defined based on several factors, such as the availability of inputs, resources, or non-availability, identifying basic needs based on customer definition and its limitations. For any outlier detection techniques, input plays a vital role. It takes a data instance or an object which has been defined through attributes. Attributes may be in the form of binary, continuous, or categorical [20]. The dataset with single attribute or multiple attributes is known as univariate or multivariate. If the dataset is multivariate, attributes may be of the same or also support different data types. After the processing of the given input, obtain the output as outliers. It may be of two categories such as labeling and scoring technique.

The labeling method label the data with regular and outlier objects. Then give the input, so that the classifier generates output with normal and outlier objects. It works like a classification algorithm. The scoring technique assigns each pattern with some outlier score, and it depends to which extent it has been considered as outliers [21]. So top most outliers are considered or cut off threshold value is fixed. Consider objects which are below the threshold value are anomalies or outliers.

In high dimensional data, search outliers in different subspaces. The benefit is, if an object found to be different in lower dimensionality, subspace gives sufficient information to what extent it is an outlier. It is applicable when an application domain has enormous dimensions. The disadvantage is that when the dimension increases, noise exists between two objects [22]. So distance measure between two objects cannot provide the exact link, and classical methods such as density or proximity relation, deteriorate when its dimensionality increases.

III. Earlier Approaches for Outlier Detection

Analysis of outliers on the text was carried out by Pawlak based on matrix factorization method. The design is in such a way that it fits into different data localities, also provides robustness when compared to existing methodologies [23]. Outlier detection using ensembles of neural networks obtained by variational approximations of the posterior in a Bayesian neural network setting. They showed that outlier detection results are better than those obtained using other efficient ensembling methods. Partitioning Around Medoids (PAM) clustering algorithm was developed to detect outliers. Objects which do not comes under cluster group forms small clusters that are outlier clusters. Remaining outliers detected by finding the absolute distance between the current cluster medoid and in the same cluster points [24]. Outlier detection based on clustering is more effective than a distance-based method. It provides accurate results, and also, improves data quality. The outlier patterns are captured and reduce the effect of outlier data in the preprocessing stage [25]. Many distance-based measures were employed to detect outliers. One among them is Manhattan distance, which outdated the performance of statistical and standard interspace measures when the value of threshold increased.

The control chart technique provides better results when compared with the linear regression method [26]. Inconsistent data can be handled with rough sets to train and test them with neural networks with the algorithm backpropagation to prevent data inconsistency [27]. The rough entropy outlier factor can detect outliers on available real-world data sets, which improves the quality of clustering with the rough set method by removing outliers. But it works only for numerical data. Rough membership function also is used to identify outliers based on two publicly available sets. For large datasets, group objects under similar features. Some objects which do not belong to any cluster are outliers.

Different kinds of labeling techniques were also used to differentiate normal and outlier objects [28]. Neural networks generally follow non-parametric and model-based methods. They are used to study the boundaries which are involved in nature. It requires both testing and training to build the data correctly. It finely tunes the network to fix threshold value and make ready for classification of data [29].

To determine outliers in the applications of knowledge discovery is more interesting than identifying inliers in a dataset. Outliers rarely perceived in data sets, described in the information table as abnormal data. In many critical applications, such as fraud detection systems, outlier detection methods are used to detect suspicious objects that may have prominent knowledge hidden in the system. The field of statistics has a long history in outlier detection, but well-known distributed data were mainly focused. So they have limitations to apply on multivariate distribution, which could be complicated to apply on real-world databases, which are extensive.

Scientists adopted a non-parametric method, and a new idea was proposed using an exemplar of distance from their neighbors, which are closer as an estimation of unusuality [30]. From audit data, detect remote outliers with computer intrusions. The distance measure is an efficient non-parametric method to identify anomalies. But the disadvantage is, it requires more time for calculation.

IV. Different Approaches to Detect Outliers

A. Statistical Technique

It is otherwise known as a model-based technique, which assumes regular data. Objects generated through statistical methods are known as ordinary objects and which do not follow this model are outliers.

1) Gaussian Model

Data points that are not in region X fit into a Gaussian distribution, in which gD for space x and function for probability density at x, denoted by gD (x). Thus, gD is used to model the normality of data. For objects in region X, we can estimate gD(y), where its probability fits into a Gaussian distribution [31]. Since gD(y) value appeared to be too low and y is obtained by the model, Gaussian is unlikely to determine an outlier. Assuming data provides a good statistic model and it holds value. It may be of parametric or non-parametric methods.

B. Parametric and Non parametric method.

Standard data objects are derived through parametric distribution with parameter Θ in the parametric method. Parametric distribution f(x, Θ) of a probability density function, provides the probability in which object x generates through the distribution. x is determined to be an outlier when the obtained value is lesser. In the non parametric approach, the statistical model has not been assumed in prior. It determines the model from the input given and does not mean that the model is always parameter-free [32]. It always makes the machine impossible to learn the model by assumption. So, the non parametric method does not consider the parameter in advance. It takes dynamically based on the parameter flexibility. Histogram and kernel density function are examples of non parametric methods.

V. Parametric Methods

A. Univariate Outlier Detection

Consider an average city's temperature value in June for the last 10 years [33]. Assume that it undergoes a distribution that is normal with two measures mean (μ) and standard deviation (σ). With maximizing the likelihood of log value, evaluate the parameters μ and σ.

$$\ln L(\mu,\sigma^2) = \sum_{i=0}^{n} \ln f(x_i \mid (\mu,\sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (1)$$

whereas n denotes the number of values taken. Derivatives concerning μ and σ^2 to be taken and solve the system by applying the first-order condition which results in the maximum likelihood as shown below:

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i \quad (2)$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3)$$

In this example, we have

$$\mu = \frac{23.0 + 27.9 + 27.9 + 28.0 + 28.1 + 28.1 + 28.2 + 28.2 + 28.3 + 28.4}{10} = 27.61$$

Temperature 23.0◦C is likely deviated 4.61◦C from the calculated mean. Region μ±3σ holds 99.7% data with assumed distribution as usual.

$$\sigma^2 = \frac{(23.0-27.61)^2 + (27.9-27.61)^2 + (27.9-27.61)^2 + (28.0-27.61)^2 + (28.1-27.61)^2 + (28.1-27.61)^2 + (28.2-27.61)^2 + (28.2-27.61)^2 + (28.3-27.61)^2 + (28.4-27.61)^2}{10} \cong 2.34$$

Hence $\hat{\sigma} = \sqrt{2.34} = 1.53$.

Apply the formula $\frac{x_i - \bar{x}}{\sigma^2}$ for individual samples to check whether the obtained value is less than or greater than 3. For example, value $\frac{4.61}{1.53}$=3.013 >3, the probability of 23.0◦C has achieved through normal distribution should be lesser than 0.15%, so it is determined to be an anomaly. Further, objects are labeled as an outlier when it is far away than 3σ by the calculated mean value of the estimated distribution, in which σ denotes standard deviation. Statistical outlier detection, which is straight forward in nature, is also be applied in visualization. The boxplot method is used to plot univariate data with the smallest nonoutlier point (Minimum), the lower quartile is Q1, Q2 is median, Q3 is upper quartile, and most significant non-outlier point (Maximum). The interquartile point (IQP) is defined as the difference between Q3 and Q1 [34]. Objects which are higher than 1.5×IQP lesser than Q1 or 1.5×IQP greater than Q3 are outliers. The region between Q1−1.5×IQP and Q3+1.5×IQP has 99.3% of entities. Figure 4 shows the boxplot method to visualize outliers.

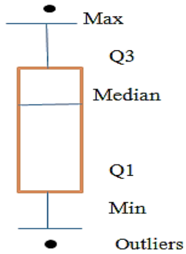

Figure 4. Outlier Visualization by Boxplot Method

The logic is the same as 3σ used as a threshold in the distribution, which is normal. An alternate approach for detecting outliers with distribution, which is normal (Grubbs test) or otherwise termed as a residual test, which is maximally normed. Z-score for each object (x) in the dataset is calculated by

$$z = \frac{|x - \bar{x}|}{s} \quad (4)$$

where mean is denoted as $\bar{x}$, and the standard deviation is denoted as s for the input. x is determined to be an anomaly if

$$Z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2{}_{\alpha/(2N),N-2}}{N-2+t^2{}_{\alpha/(2N),N-2}}} \quad (5)$$

where the value is taken by t-distribution as t2α/(2N), N−2 at the intent level of α/(2N), represents the total objects in the dataset as N.

B. Multivariate Outlier Detection

Data with two attributes or variables more than that are known as multivariate data. Many univariate techniques have transformed to multivariate data [35]. The primary goal of multivariate data is transformed into univariate data. Some of the examples are discussed below:

1) Mahalanobis distance

Let us consider a data set that is multivariate, where the mean vector is $\bar{o}$. For an object, o, Mahalanobis distance from o to $\bar{o}$ is calculated as

$$MDist(0, \bar{o}) = (0 - \bar{o})^T S^{-1} (0 - \bar{o}) \quad (6)$$

where the covariance matrix is S, univariate variable is $MDist(0, \bar{o})$, and then observe Grubb's test to this quantity. So, transforming of multivariate data is as follows:

1. For the data set, which is multivariate, then calculate the mean vector.
2. Calculate $MDist(0, \bar{o})$, for each object o, and then Mahalanobis distance from o to $\bar{o}$.
3. For univariate data, outliers are to be detected $MDist(0, \bar{o}) |\ 0 \in D$.
4. $MDist(0, \bar{o})$ detects outlier, then o is an outlier.

The next example shows χ^2-statistic, which is used to calculate the distance between an object and to the mean value.

2) χ^2-Statistic Method

The multivariate outliers are detected using χ^2-statistic [36] under the suspicion of the normal distributed method. The χ^2-statistic for each object, o, is defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i} \quad (7)$$

where o_i denotes the i^{th} dimension of object o, mean is represented by E_i, and n represents its dimensionality. The entity is identified as an anomaly when its χ^2-statistic is large.

C. Combined Parametric Distributions

Data generated through normal distribution are suits well in many situations. The main disadvantage of this method is that it won't work for high dimensional data. At that time, use a mixture of parametric distributions.

1) Multiple Parametric Distributions

For instance, C_1 and C_2 are two large clusters. Applying normal distribution over the data cannot be worked out here because the intended mean is not located within any cluster rather than between two clusters [37]. Objects which are lying in between the clusters cannot be determined as an outlier even though they are very closest to the mean. This problem has been overcome by assuming standard data into multiple distributions. Let us take any object, o, in the dataset, their normal distribution is $\Theta_1(\mu_1, \sigma_1)$ and $\Theta_2(\mu_2, \sigma_2)$, then for any object *o*, combined distributions are given by

$$Pr(0|\Theta_1, \Theta_2) = f_{\Theta_1}(0) + f_{\Theta_2}(0) \quad (8)$$

where probability density function is denoted as f_1 and f_2, Θ_1 and Θ_2, respectively, to know about the parameters such as μ_1, σ_1, μ_2, σ_2 in a dataset, expectation-maximization (EM) algorithm is used to perform clustering. Each cluster represents a known normal distribution. If an object, o, does not belong to any cluster, which means, the obtained probability is very slow when combined with the two distributions.

2) Multiple Clusters

Data objects may lie either in C_1 or C_2. Objects which do not lie in between C_1 or C_2 are noisy data that are distributed equally in the space [38]. If there exists a small cluster C_3, which is not very much close to any of the clusters C_1 or C_2 is highly suspected. So, objects in C_3 are determined as outliers but determining C_3 as an outlier is a tedious one, even though it follows normal or multiple distributions. Because in C_3, the probability of objects is higher due to noisy objects, as o, since they have a local density value, which is higher in C_3. Outliers are represented in a broader area if the distribution has a significant variance. The representation is σ outlier=kσ, where the parameter gets from the user is denoted as *k*, and the standard deviation is σ, for regular distribution to generate data as usual. Indeed, we can use the EM algorithm to know about the parameters.

VI. Non parametric Methods

In non-parametric methods, normal data are not assumed in prior. It makes only lesser assumptions about the data, so it applies to many domains.

A. Histogram Approach

Consider an electronic shop, which calculates the customer purchase amount for an individual transaction. Figure 5 shows the graph-based on percentages for recorded transactions. Most of the transactions, like sixty percentage, are lie between Rs0 and Rs1000 [39]. So, histogram is used for nonparametric statistical models to detect outliers. It is a frequently used nonparametric statistical model to identify outliers. It involves two steps:

1) Construction of Histogram

Based on inputs, construct the histogram. It may be univariate or multivariate based on multidimensional input. It requires parameters from the user, not from any prior statistical model. The construction of a right histogram requires type, which has specified like equal width or depth [40], several bins needed, or its fixed size. The parameters need not specify data distribution like Gaussian.

We can identify outliers quickly for every individual object after the construction of a histogram. Objects which fall into the histogram bin's are regular; otherwise, they are anomalies. For the histogram, assign outlier rank to objects. The outlier rank of an entity depends upon the bin capacity, where the object belongs. Outlier rank of an entity depends upon the bin capacity, where the object belongs. It can be calculated as 1-(60%+20%+10%+6.7%+3.1%) results in transactions of 0.2% for amount greater than Rs5000.

Consider 60% of transactions are between Rs0 and Rs1000, 20% of transactions are between Rs1000 and Rs2000, 10% of transactions are between Rs2000 and Rs3000, 6.7% of transactions are between Rs3000 and Rs4000, and 3.1% of transactions are between Rs4000 and Rs5000.The outlier score is generated by the inverse of the bin volume where the object falls. However, outlier score of amount Rs7500 is $\frac{1}{0.2\%}$ =500, and for Rs385 is $\frac{1}{60\%}$ =1.67. From the score, Rs7500 is determined to be an outlier rather than Rs385 because it falls into the bin of 60% transactions. Figure 5 shows the histogram for customer purchase transactions.

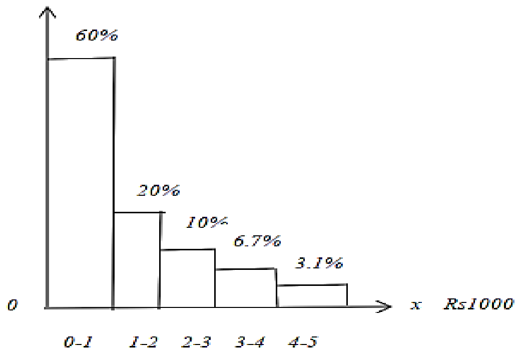

Figure 5. Histogram for customer purchase transactions

The drawback of using a histogram is choosing the size of the bin. If the capacity of the bin is too low, objects(regular) can have a void or infrequent bin, so there is a chance to misidentify regular objects as outliers [41]. Suppose if the bin capacity is high, many outlier objects have mislabeled as ordinary objects, and they occupy bins. It results in false-positive rate with low precision value. To overcome the problem, introduced the concept of kernel density-based method.

B. Kernel Density - Based

Consider the observed object as an indicator, from which high probability function has been determined by its surrounding region [42]. The distance between a particular point and observed object is considered as the high probability density function. A kernel function is used to estimate some points with its neighbor. The K() method is an integral function with positive real-valued, shows below :

$$\int_{-\infty}^{+\infty} \mathrm{K(u)du} = 1 \tag{9}$$

$$K(-u)=K(u) \tag{10}$$

For all u values. The standard Gaussian function is a frequently used kernel function with 0 mean value and 1 variance value.

$$K\left(\frac{\mathrm{x-x_i}}{\mathrm{h}}\right)= \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-x_i)^2}{2h^2}} \tag{11}$$

A random variable f has objects x1….xn, which are independent and equally distributed. The probability density function for kernel density function is as follows:

$$\hat{\mathrm{f}}_h\,(x)=\frac{1}{\mathrm{nh}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{k}\frac{(\mathrm{x-x_i})}{\mathrm{h}} \tag{12}$$

where kernel function represents k(), and h represents bandwidth. Once approximate the dataset with probability density function with a kernel density estimated approach, then the estimated density function $\hat{\mathrm{f}}$ is used to detect outliers. Take an object, o, its probability is estimated through $\hat{\mathrm{f}}$(o), and the stochastic process is used to generate the object. If the value of f(o) is high, then it is determined as natural objects and otherwise labelled as an outlier. In general, statistical methods are used to distinguish between normal and outlier objects. Only data that satisfies the constraints, in statistical methods are justifiable. For high dimensional data, distribution is still challenging [43].

Compute cost based on the models. The simple parametric model based on linear time typically fits the parameters. But for complex models, the EM algorithm has used with several iterations. The iterations are linear to dataset volume. The learning model for kernel density-based estimation is cost-effective. For learning, the cost of detecting outliers is very less per object.

VII. Approaches Based on Proximity

The similarity between objects is obtained through distance measure in feature space. Objects which are farther from its measure are outliers, and it is of two types, such as distance and density-based approaches [44]. The distance-based method considers its neighbors along with its radius. If its neighbor does not have enough points, then the object is declared as an outlier. Density-based methods identify the object density with its neighbors. If an object has a lower density value when compared with its neighbors, are identified as outliers.

A. Distance - Based Method

For a set of objects, D, a threshold value, r is determined. Then for each object, o, the r-neighborhood can be calculated [45]. If the objects present in D are farther from o or does not exist in the r-neighbor of o, then outliers are identified. Consider r, (r≥0) be a threshold for a distance, and π ($0 < \pi \leq 1$) be a fractional threshold. An object, o, is a DB(r,π)-outlier if

$$\frac{\|(\mathrm{o}'|\mathrm{dist}(\mathrm{o},\mathrm{o}')\leq \mathrm{r}\|}{\|\mathrm{D}\|} \leq \pi \tag{13}$$

where dist(·,·) is a distance measure. Figure 6 shows nested loop algorithm.

1) Algorithm

```
Input:  tuples D, r is threshold (r > 0) and
        fractional threshold π (0 < π ≤1);
Output: DB (r, π) outliers in D.
Method:
  For i=1 to p do
    Count ⟵0
    For j=1 to p do
      If i≠j and dist(oi, oj)≤ r Then
        ct ⟵ct+1
        If ct ≥π •p Then
          Exit{oi is not a DB(r,π) outlier}
        End if
      End if
  End for
  Print oi{oi is a DB(r,π) outlier based on equation (13).
End for;
```

Figure 6. Nested loop algorithm

The nested loop approach takes the time of $O(n_2)$. The CPU time is linear to dataset capacity. If the dataset holds a small number of outliers, iteration stops earlier. So, a limited fraction of the dataset is evaluated. If the dataset is large, it does not fit into the main memory, where constructing the cost for the nested loop is high.

Consider m pages in main memory, then *m*-1 pages are used to hold objects, and the last page runs the inner loop. The inner loop has not stopped until running all the pages, which sometimes happens. Accordingly, the algorithm has input/output (I/O) cost as O; b is the number of objects, which is on a single page. The cost factor of the nested approach is high, based on two aspects. The whole dataset has to be inspected to check outliers. It has been improved by identifying its neighbors which are close to it. Second, the objects are checked one by one in the nested loop method. Grouping objects can improve it according to proximity relation so that the outliers can be detected group by group rather than individually.

B.Grid-Based Approach

Outlier detection based on distance approach uses CELL, a grid - based method in which space for dataset has been partitioned to the multidimensional grid [46]. Each cell is hypercube with its diagonal length r_2 where r denotes threshold. For a single dimension, the length of each cell is $\frac{r}{2\sqrt{1}}$.

2	2	2	2	2	2	2
2	**2**	**2**	2	2	2	2
2	**2**	**1**	1	1	2	2
2	2	1	C	1	2	2
2	2	1	1	1	2	2
2	2	2	2	2	2	2
2	2	2	2	2	2	2

Table 2. Cell Method.

Table 2 represents a 2D dataset. The length of each cell is $\frac{r}{2\sqrt{2}}$. Now, if we take C, cells immediate to C is level 1 cells other than that are level 2 cells. It has some properties which are as follows:

1) Cell Property of Level-1

Take any point as x and possible point as y of C, and then their distance is dist(x,y)≤r.

2) Cell Property of Level-2

Any possible point as x and any point as y of C, then their distance is dist(x,y)≥r, then the level-2 cell has y. Let us consider (a,b_1,b_2) where a be the number of objects in the cell C, $b_1$is the objects in level-1, and b_2 be the objects in level 2. The following rules can be applied.

3) Pruning Rule of Level - 1 Cell

If a + b1> dπne based on level 1 cell property, then objects present in C are not DB(r,π)-outliers. Because of objects in C and level 1 cells in r-neighbor of o and atleast neighbors as $[\pi, n]$.

4) Pruning Rule of Level -2 Cell

If a + b1 + b2<dπne+1 based on level 2 cell property, then all objects in C are DB(r,π)-outliers. Because each r neighbors have less than dπne, other objects.

C.Density-Based Outlier Detection

The DB(r,π)-outliers is a type of distance-based outlier detection method. It gives the global view of the dataset. Outlier detected in this method are known as global outliers. If an object, o, is quantified with the parameter r, is atleast (1-π) ×100% of objects in a dataset. To fix outliers, we need two parameters, such as r and π. But complex structure need, outliers can be fixed based on their local neighborhoods rather than global distribution [47].

D.Class Outlier Factor

In class outlier factor, objects in a dataset are given rank with parameters Q, which are top class outliers and P be the nearest neighbors. The rank of each object has obtained through the formula:

$$\begin{aligned} ClsOutFac &= ProbClsLbl(O,P) \\ normal&\left(\begin{array}{c} deviation(O) + \\ normal\big(PDistance(O)\big)\end{array}\right) \end{aligned} \tag{14}$$

where *ProbClsLbl(O, P)* represents the probability of class label of object *O* for its nearest neighbors *P*. *normal(deviation(O))* and *normal(PDistance(O))* denotes the values which are normalized, and their range falls between 0 to 1 [48]. *deviation(O)* denotes how much an object deviates from other objects within the same class. *PDistance(O)* denotes the summation value of the distance between an object *O* and its neighbors *P*. It also adds a Boolean attribute outlier, which identifies outliers when it is assigned real value. Also, a unique attribute named class outlier factor to determine the degree of outlier classes.

E.Local Outlier Factor

It is also known as outlier detection based on density method according to [49]. Its *P* neighbors can identify the locality of an object with its estimated density measured. Each local object density has compared with its *P* neighbor's local density value.

Declare the lesser values are outliers. A point which is nearest from neighbor to reach its destiny is called reachable distance. The outlier detection method measures the object density with its neighbors. Calculate the local outlier factor by the average ratio of local reachable distance *P* nearest neighbors.

VIII. Methods Based on Clustering

Clustering techniques detect outliers based on the relationship between clusters and objects. Since outlier objects may belong to very small or far away clusters and sometimes it does not belong to any cluster. So, the representation of outliers is significantly related to data clusters.

A.Density-Based Approach.

Gregarious animals such as goats and deers lived together and migrated in flocks. Animals that are not part of flock or animals which do not belong to any cluster are outliers. Such animals might be either wounded or lost [50].

Figure 7 shows that animals are living in groups. By using DB-SCAN, which is a density-based clustering approach, black points represent the objects which are in clusters and white point, x, which does not cover any cluster or different from other cluster behavior is an outlier. Another alternative approach is considering the distance from any object to its nearest cluster. When its distance is measurably significant, objects are outliers. Individual outliers in a dataset can be detected based on this method.

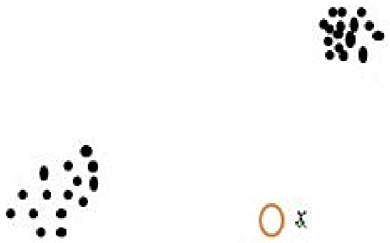

Figure 7. Density-Based Outlier Detection

B.Distance Based Approach

Data forms three clusters based on k-means algorithm, as shown in Figure 8, with different symbols. The + symbol is used to denote the center of each cluster. Assign outlier scores to each object, o, based on their distance between the object and its center, which is close to them[51].

Let co is the center which is nearer to o; then dist(o, co) is the distance between o and co , and their average distance is l_{c_o} concerning co and objects assigned with o. The $\frac{\mathrm{dist}(o,c_o)}{l_{c_o}}$ ratio measure, shows dist(o,c_o) deviate from its average value. If the ratio is significant, the object o is farther away from the center, so that o is an outlier. In Figure 8, points x, y, and z are farther away for their centers, are detected being of outliers.

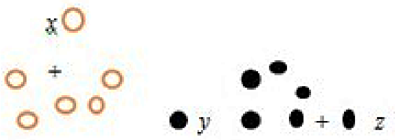

Figure 8. Distance-based clustering to the object and their centers

IX. Rough set Theory

It was developed in the early 1980s by Zdzislaw Pawlak [35]. It is a potent mathematical tool for dealing with inaccurate decision-making situations. Incomplete information processing is based on the concept of approximation for two crisp sets of approximations. Rough sets do not need any preliminary or additional information for processing. Topological, indoor, and closure operations define rough sets called approximations [52]. The motivation for rough set theory is to represent the universe in terms of the equivalence relation.

With rough membership also outliers can be detected to demonstrate two data sets available publicly. There is an active link between vagueness and uncertainty. Vagueness is associated with sets (concepts), while uncertainty is associated with sets of elements. The rough set approach shows a clear link between uncertainty and vagueness.

A.Rough Classification

The boundary region or boundary line segregates approximation, which is lower from upper value [53]. Either lower or upper approximated levels do not cover those certain elements. Figure 9 shows the rough set classification system's information table, indiscernibility, attribute reduct and core method, partitioning and generation of rules.

The information table is also known as a decision table that contains objects (tuples) and attributes(fields) with non-empty data. Also, it might include conditional and decisional attributes.

To avoid redundancy or repetition of data, indiscernible or similar values are grouped based on their characteristics or behavior. Any union of elements in a dataset defines a crisp set, whereas the union set sometimes does not occur as a crisp set, then it is said to be rough or vague.

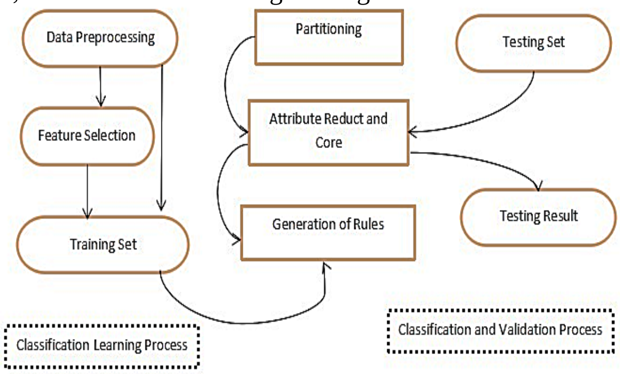

Figure 9. Roughset Classification

Some of the attributes which removed from the dataset without affecting its essential properties are known as attribute reduction. The core is an intersection of all reducts, which removes a subset of prominent elements.

Partitioning provides classification with high quality. It transforms the continuous value into discrete or intervals to group attributes. Always cut associates with discriminant as a pair (p,q) where p represents a continuous variable and q is cut value used to separate two disjoint subintervals.

Rules have been generated when conditional attributes are satisfied, then decision attributes are executed and represented as IF and Then Decision.

B.Rough Clustering

It is also similar to rough set theory based on lower and upper approximations. The lower approximation of rough clustering indicates that the object should belong to that particular cluster only, and the upper approximation of a rough clustering indicates that objects can be a member of other clusters also [54].

The clustering algorithm should follow the properties of rough sets as object should be a member of one lower approximation at the most, and intra clusters can have objects in upper and lower approximation also.

1) Improved Rough k-means Clustering

The selection of random centroids results in weak clustering. Centroids are defined prior and classified into three types, such as similarity-based, entropy-based, and dissimilar to similar based.

(a) Similarity-Based Measure

Between objects, similarity measure is calculated, and select the highest similarity value object as centroid (prior) for doing the clustering process [55]. The algorithm shows below:
Input: Lower and Upper Approximation Values
Output: Grouped objects
Step 1: The maximum similarity value object has been placed in a lower approximation level and also by second property in can also be a member of upper approximation.

$$Sim_{mn} = \mathrm{e}^{-\alpha D_{mn}} \tag{15}$$

where α is a constant and D_{mn}denotes the distance between two objects which can be obtained by the formula

$$D_{mn} = \sqrt{\sum_{k=1}^{l}(X_{mk} - X_{nk})} \tag{16}$$

The D_{mn} values should lie between 0.0 and 1.0. When the distance between objects and their mean distance is equal, then find α by fixing similarity measure value as 0.5.

$$\alpha = -\frac{ln\,0.5}{\overline{D}} \tag{17}$$

where $\overline{D}$ represents objects average distance which has obtained with the formula

$$\overline{\mathrm{D}} = \frac{1}{\mathrm{P}}\sum_{\mathrm{r}=1}^{\mathrm{P}}\sum_{\mathrm{n}>\mathrm{m}}^{\mathrm{P}} \mathrm{D}_{\mathrm{mn}} \tag{18}$$

$$\mathrm{TotSim}_{\mathrm{m}} = \sum_{\mathrm{n}\in\mathrm{x}}^{\mathrm{n}\neq\mathrm{m}} \mathrm{Sim}_{\mathrm{mn}} \tag{19}$$

Then $TotSim_m$ is a total similarity measure between objects. The object which holds a more significant similarity measure is fixed to be a centroid (prior) for successive clusters.
Step 2: New mean value has obtained by using the formula

$$mean_t = \begin{cases} Z_{lr}\sum_{Y_{t\in D_t}}\frac{T_s}{|D_t|} + Z_{BND}\sum_{Y}\frac{T_s}{|D_t^B|} \; for\; D_t^B \neq 0 \\ Z_{lr}\sum_{Y_{t\in D_t}}\frac{T_s}{|D_t|} \;\; otherwise \end{cases} \tag{20}$$

The values Z_{lr} and Z_{BND} represent the lower and boundary conditions for the objects for a rough cluster, whereas $|D_t|$ represents the total objects present in the dataset.The boundary condition is $\left|\overline{D_t} - \underline{D_t}\right|$
Step 3: For an object T_s, calculate the closest mean $\mathrm{cm}_\mathrm{h}$by using the formula

$$cd_{s,h}^{min} = min_{t=1\ldots t}\, cd(T_s, cm_t) \tag{21}$$

where T_s is an upper approximation which belongs to cluster h.
Step 4: Identify the mean which is closer by fixing the threshold value as

$$B = \{b: cd(T_s,\mathrm{cm_t}) - (\, cd(T_s,cm_h) \leq \in \cap\; h \neq t \tag{22}$$

Step 5: Continue Step 2 to satisfy the condition or else stop the process.

(b) Entropy Measure

Calculate entropy measure between objects and an object which has a lower entropy measure is selected as centroid (prior) for clustering [56]. The algorithm is as follows:
Input: Lower and Upper Approximation Values
Output: Clustered objects in lower and upper approximations.
Step 1: Objects distributed in the lower and upper approximation of the same cluster those are having minimal entropy measure. The following formula can calculate it.

$$Etpy_d = -\sum_{d\in x}^{d\neq c}((Sim_{mn}\, log_2 Sim_{mn}) + (1 - Sim_{mn}) \;\; log_2(1 - Sim_{mn})) \tag{23}$$

where $Etpy_d$ denotes the entropy measure of an individual object, and $\mathrm{Sim_{mn}}$ denotes the similarity measure between objects, and then they are added to successive clusters.
Step 2: From the obtained centroid, calculate the new mean according to equation 19.
Step 3: Obtain the closest mean according to equation 20.
Step 4: Fix the threshold by using equation 21.
Step 5: Continue Step 2 to satisfy the condition or else stop the process.

(c) Dissimilar to Similar Based

Objects which are having minimal dissimilar and similar values [57] are selected as centroid(prior) for clustering. The algorithm is as follows:
Input: Lower and Upper Approximation Values
Output: Clustered objects in lower and upper approximations.
Step 1: The proportional value of dissimilar to similar value should be minimum in each of the clusters derived from

$$DisSim_d = \sum_{c\in n}^{c=d}\frac{(1-Sim_{mn})}{Sim_{mn}} \tag{24}$$

where $DisSim_d$denotes the ratio of dissimilarity to similarity values for an object.
Step 2: From the obtained centroid, calculate the new mean according to equation 19.
Step 3: Obtain the closest mean according to equation 20.
Step 4: Fix the threshold by using equation 21.
Step 5: Continue Step 2 to satisfy the condition or else stop the process.

C.Outlier Detection with Roughsets

Objects which differ from their behavior or characteristics are known as outliers. Such outliers are identified with rough sets.

Entropy measure is used to define the uncertain properties of an object. So, it can be combined with rough sets to determine outlier factor of an object which shows below:

1)Outlier Detection-Numerical Data

The outlier factor based on rough entropy measure is used to find an outlier degree for every object present in the universe [58]. Let the information table represented as *IT*= (*W*,*F*) where *W* denotes universe and F denotes attributes. The outlier factor based on rough entropy measure for an individual object is as follows:

$$Etrpy_k = Q_k * \log_2 Q_k \tag{25}$$

where Q_k is the measured distance between the object and centroid.

$$EBROF_k = \left(\frac{(Etrpy_k^{max} - Etrpy_k^{min})}{2} * \left(1 - \frac{|Ct_k|}{n}\right)\right) \tag{26}$$

where $Etrpy_k^{max}$ and $Etrpy_k^{min}$ denote the maximum and minimum entropy measure of objects in the k th cluster. If $Etrpy_k < EBROF_k$, then that object is identified as an outlier.

2) Outlier Detection- Categorical Data

The rare events have measured by its average density of each object based on its equivalence classes [59]. The definition of a dataset is a collection of attributes and objects arranged in order.

Step 1: Measure the rare events by its average density of each object based on its equivalence classes. The definition of a dataset is a collection of attributes and objects arranged in order.

$$Ind(T) = \left\{ \begin{matrix} (a, b \in W \times W \mid \forall\, c \in T, f(a,c) \\ = f(b,c) \end{matrix} \right\} \tag{27}$$

Step 2: The average density of attributes can be calculated by the formula as follows:

$$AvgDens(a) = \frac{\sum_{c \in C} AvgDens_c(a)}{|C|} \tag{28}$$

$$AvgDens_c(a) = \frac{|[a]_{\{c\}}|}{|W|} \tag{29}$$

Step 3: Complement Entropy for an individual object can be determined as follows:

$$CmpEntrpy = \sum_{j=1}^{m} \frac{|A_i|}{|W|}\left(1 - \frac{|A_i|}{|W|}\right) \tag{30}$$

Step 4: The weighted density of each object is as follows:

$$WgtDens(a) = \sum_{c \in C} AvgDens_c(a) * Wgt(\{c\}) \tag{31}$$

$$Wgt(\{c\}) = \frac{1 - Entpy(c)}{\sum_{l \in C}(1 - Entpy(l))} \tag{32}$$

Step 5: From this, fix the threshold value. Objects which are lesser than the threshold values are outliers.

D.Experimental Analysis

To show the performance of rough entropy-based weighted density method (REBWDM), it has been compared with existing outlier detection algorithms such as Local Outlier Factor (LOF), Feature Bagging (FB), Histogram Based Outlier Sequence (HBOS), Isolation Forest (IF), K Nearest Neighbour (KNN) and Average KNN. Wine dataset from UCI repository is taken for analysis with 178 objects and 14 attributes. Local outlier factor determines the neighborhood distance by estimating its density. The similar density values form a group, and substantial lesser values are outliers. Feature Bagging splits the dataset into sub-samples, and it fixes the base estimators [59]. The local outlier factor is a default base estimator. Acquire prediction accuracy by combining or averaging all estimators. Histogram based outlier sequence is an unsupervised method, which detects outliers based on histograms. Isolation forest suitably fits for multidimensional data.

Dataset is partitioned into a set of trees, whereas the isolated points are outliers. K nearest neighbor algorithm is mainly used in the problem of classification and regression. Similarity based on distance measures calculates the neighbor's majority vote.

Average KNN creates a super sample for each class by training its samples, and finding the average Rough entropy-based weighted density algorithm determines weighted density value for all objects and attributes by considering its indiscernible relation, complement entropy and an average weight of attributes and objects. Figure 10 shows the comparison chart for roughset based entropy measure weighted density over existing methods.

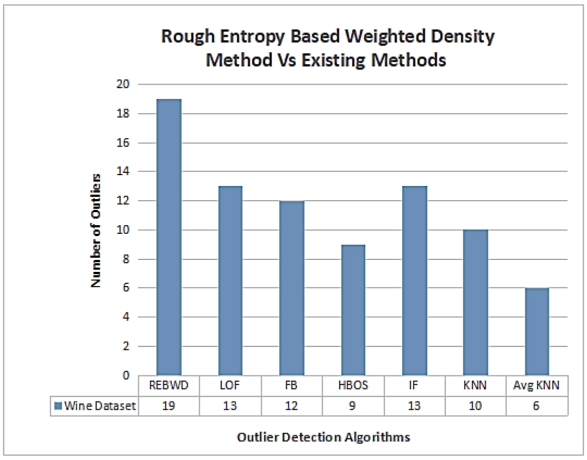

Figure 10. Comparison chart for roughset based entropy measure weighted density method with existing methods

E.Performance Evaluation

The performance of roughset based entropy measure weighted density and existing methods is compared by calculating its measures such as accuracy, specificity, sensitivity, precision, recall, and F1 score [60]. The classifier provides true positive (tp), true negative (tn), false positive(fp) and false negative(fn) values. The percentage of testing data that are classified by the classifier provides accuracy.

$$Accuracy = \frac{tp + tn}{tp + fn + tn + fp} \tag{33}$$

The proportion of true negative and true positive values that are identified by the classifier denotes specificity and sensitivity or recall.

$$Specificity = \frac{tn}{tn+fp} \tag{34}$$

$$Sensitivity = \frac{tp}{tp+fn} \tag{35}$$

$$Recall = \frac{tp}{tp+fn} \tag{36}$$

Precision measures the proportion of instances which are relevant to the retrieved instances.

$$Precision = \frac{tp}{tp+fp} \tag{37}$$

F1 score considers both false positive and false negative values, so it determines the average weight of both precision and recall. It provides better results than accuracy.

$$F1\ score = \frac{2*precision*recall}{precision+recall} \tag{38}$$

Measures	**REBWDM**	**LOF**
Accuracy	95.23	92.50
Specificity	94.44	94.11
Sensitivity	95.87	91.30
Precision	95.81	95.45
Recall	94.73	91.30
F1 Score	95.27	93.32

Table 3. Performance Measures of Wine Dataset.

Table 3 represents the performance evaluation of the wine dataset with rough entropy-based weighted density and existing local outlier factor method.

F.Benefits

This survey article facilitates the readers to understand the concepts of various outlier detection techniques. A detailed literature review provides them a comprehensive layout. From this, know the limitations of each algorithm and applied to different domains. The working flow of different parametric and non-parametric outlier detection methods has discussed. Rough sets based on classification, clustering and outlier detection method for categorical and numerical data have discussed. The literature review would be helpful for researchers to continue their research with data mining along with rough sets.

X. Conclusion

The availability of data in various fields cannot be used directly in real-time applications. There exist missing or null values, which are not well formulated. Some objects which deviate from other objects based on behavior or characteristics are known as outliers. Different methods of outlier detection are discussed and proposed a detailed literature survey. Several outlier detection methods are specified, which maps to a single application domain. Rough sets are handled vagueness and uncertainty of data present in the application domains. The concepts of rough classification, rough clustering, and different outlier detection methods are provided with algorithms and solved with suitable examples. This survey provides an overview of outliers and existing outliers by classifying them into different dimensions. Wine dataset from the UCI repository has been taken to prove the performance of roughset based entropy measure weighted density method over existing methods. However, roughset based entropy measure weighted density method provides a solution for single granulation sets. It may be extended to detect outliers in two universal sets, multi granulation sets, neutrosophic sets and for dynamic inputs also. Hopefully, this survey helps researchers to provide a different perspective on outlier detection methods.

References

[1] E. Achter, H.P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, A. Zimek. “Visual Evaluation of Outlier Detection Models”. *In Proc. International Conference on Database Systems for Advanced Applications (DASFAA),* pp. 396-399, 2010.

[2] C.C. Aggarwal, P.S. Yu. “Outlier detection for high dimensional data”. in *Proc. ACM-SIGMOD, Int. Conf. Management of Data (SIGMOD’01), Santa Barbara, CA*, pp. 37-46, 2001.

[3] A. Arning, R. Agrawal, P. Raghavan. “A linear method for deviation detection in large databases”. in *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD),* Portland, pp. 1-15, 1996.

[4] P. Ashok, G.M Kadhar Nawaz. “Outlier Detection Method on UCI Repository Dataset by Entropy Based Rough K-means”, Journal of Defence Science, 11 (1), pp. 113-121, 2016.

[5] V. Barnett, T. Lewis. *Outliers in statistical data*. John Wiley and sons, 1994.

[6] A. Geetha Mary. “Evaluation of Features to Identify a Phishing Website Using Data Analysis Techniques.” in *Information Systems Design and Intelligent Applications. Springer*, *Singapore*, pp. 267-276, 2019.

[7] S.D. Bay, M. Schwabacher. “Mining distance-based outliers in near linear time with randomization and a simple pruning rule”. in *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC,* pp. 1-15, 2003.

[8] R.J. Beckman, R.D. Cook. “Outlier ……… S”, *Technometrics*, 25(2), pp:119-149, 1983.

[9] M.M. Breunig, H.P. Kriegel, J. Sander. “Identifying density based local outliers”, in *Proc Acm Sigmod Conference*, pp. 93–104, 2000.

[10] V. Chandola, A. Banerjee, V. Kumar. “Anomaly Detection – A Survey”, *ACM Computing Surveys,* 41(1), pp.58-66, 2011.

[11] R. Chitrakar, H. Chuanhe. “Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification.” in *Proceedings of 8th IEEE International Conference on Wireless Communications, Networking and Mobile Computing* (WiCOM), pp. 1-5, 2012.

[12] R. Chitrakar, H. Chuanhe. "Anomaly detection using Support Vector Machine classification with k-Medoids clustering." in *Proceedings of IEEE Third Asian Himalayas International Conference on Internet (AH-ICI)*, pp. 1-5, 2012.

[13] A. Christy, G. Meera Gandhi, S. Vaithyasubramaniyan. "Cluster based outlier detection algorithm for healthcare data", *Procedia Computer Science*, 5(5), pp: 363-387, 2012.

[14] D. Dasgupta, F. Nino. "A comparison of negative and positive selection algorithms in novel pattern detection." in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 125-130, 2000.

[15] M. Ester, H.P. Kriegel, J. Sander, X.A. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pp. 1-6, 1996.

[16] F. Jiang, Y. Sui, C. Cao. "Outlier Detection Based on Rough Membership Function, Rough Sets and Current Trends in Computing." in *International Conference on Rough Sets and Current Trends in Computing, pp. 388-397. Springer, Berlin, Heidelberg*, pp. 388-397, 2006.

[17] A. Ferone, A. Maratea. "Rough Graded Possibilistic Meta: Outlier Detection in Granular Clustering." in *Proceedings of the 20th International Conference on Computer Systems and Technologies*, pp. 105-109, 2019.

[18] A. Ghoting, S. Parthasarathy, M. Otey. "Fast mining of distance-based outliers in high dimensional spaces." in *Proc SIAM Int Conf on Data Mining (SDM) dimensional spaces, Bethesda, ML*, pp. 1-6, 2006.

[19] F.E. Grubbs. "Procedures for detecting outlying observations in samples", *Technometrics*, 11(1), pp. 19-21, 1969.

[20] G. Liu. "Rough set theory based on two universal sets and its applications", *Journal of Science Direct*, 23 (1), pp. 110-115, 2010.

[21] G. Lin, Y. Qian, J. Li. "NMGRS: Neighborhood-based multi granulation rough sets", *International Journal of Approximate Reasoning*, 53(7), pp. 1080-1093, 2012.

[22] D. Hawkins. "Identification of outliers", *Monographs on Applied Probability and Statistics*, pp. 1-20, 1980.

[23] V. Hodge, J. Austin. "A survey of outlier detection methodologies", *Artificial Intelligence Review*, 22(2), pp. 85-126, 2004.

[24] J. Han, M. Kamber, J. Pei. *Data Mining concepts and techniques*. Elsevier, 2012.

[25] J. Liang, F. Wang, C. Dang, Y. Qian. "An efficient rough feature selection algorithm with a multi-granulation view", *International Journal of Approximate Reasoning*, 53(6), pp. 912-926, 2012.

[26] E.M. Knorr, R.T. Ng. "A unified approach for mining outliers." in *Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada*, pp. 1-8, 1997.

[27] E.M. Knorr, R.T. Ng. "Algorithms for mining distance-based outliers in large datasets." in *Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY*, 1998.

[28] Kovarasan, R. Kambattan, M. Rajkumar. "An Effective Intrusion Detection System Using Flawless Feature Selection, Outlier Detection and Classification." in *Progress in Advanced Computing and Intelligent Engineering*. *Springer, Singapore*, pp. 203-213, 2019.

[29] J. Komorowski, Z. Pawlak, L. Polkowski, A. Skowron. "Rough sets: A tutorial", *Rough fuzzy hybridization: A new trend in decision-making*, pp. 3-9,1999

[30] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava. "A comparative study of anomaly detection schemes in network intrusion detection." in *Proceedings of the Third SIAM International Conference on Data Mining. SIAM*, 2003.

[31] M. Markou, S. Singh. "Novelty detection: A Review-Part 1: Statistical Approaches", *Signal Processing*, 83(12), pp. 2481-2497, 2003.

[32] M. Markou, S. Singh. "Novelty detection: A Review-Part 2: Neural Network based approaches", *Signal Processing*, 83(12), pp. 2499-2521, 2003.

[33] A. McCallum, K. Nigam, L.H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching." in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA*, 2000.

[34] A. Patcha, J.M. Park. "An overview of anomaly detection techniques: Existing solutions and latest technological trends", *Computer Networks*, 51(12), pp. 3448-3470, 2007.

[35] Z. Pawlak. "Rough Sets", *Journal Computer and Information Sciences*, 11(1), pp. 341-356, 1982.

[36] S. Peddabachigari, A. Abraham, C. Grosan, J. Thoms. "Modeling intrusion detection system using hybrid intelligent systems", *Journal of network and computer applications*, 30(1), pp. 114-132, 2007.

[37] M.I. Petrovskiy. "Outlier detection algorithms in data mining systems", *Programming and Computer Software*, 29(4), pp. 228-237, 2003.

[38] F. Preparata, M. Shamos. *Computational Geometry: an Introduction*. Springer Verlag, 1988.

[39] Y.H. Qian, J.Y. Liang, Y.Y. Yao. "MGRS: A multi-granulation rough set", *Information Sciences*, 180(6), pp. 949-970, 2010.

[40] R. Banshal, N. Gaur, S.N. Singh. "Outlier Detection-Applications and Techniques in Data Mining", *IEEE Conference*, 44(12), pp. 2862-2870, 2016.

[41] R. Kannan, H. Woo, C.C. Aggarwal. "Outlier Detection for text data-An extended version", arXiv preprint arXiv:1701.01325, 23(1), pp. 61-69, 2000.

[42] R. Kaur, S. Singh. "A review of social network centric anomaly detection techniques", International Journal of Communication Networks and Distributed Systems, 23(1), pp. 61-69, 2016.
[43] P.J. Rousseeuw, A.M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA, 1987.
[44] I. Ruts, P.J. Rousseeuw. "Computing depth contours of bivariate point clouds", *Computational Statistics and Data Analysis*, 23(1), 153-168, 1996.
[45] S. Senthil Kumar, H. Hannah Inbaran. "Optimistic Multi-granulation Rough Set Based Classification for Medical Diagnosis", in *Procedia Computer Science*, 47(1), pp. 374-382, 2015.
[46] T. Sangeetha, A. Geetha Mary. "A Rough Entropy-Based Weighted Density Outlier Detection Method for Two Universal Sets." in *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*. *Springer, Singapore*, pp. 509-516, 2019.
[47] N.N.R. Ranga Suri, G. Athithan, "Outlier Detection." in *Outlier Detection: Techniques and Applications*. Springer, Cham, pp. 13-27, 2019.
[48] J. Tang, Z. Chen, A.W. Fu, D.W. Cheung. "Capabilities of outlier detection schemes in large datasets, framework and methodologies", *Knowledge and Information Systems*, 11(1), pp. 45-84, 2006.
[49] D.H. Tang, Z. Cao. "Machine Learning-based Intrusion Detection Algorithms", *Journal of Computational Information Systems*, 5 (6), pp. 1825-1831, 2009.
[50] D. Savage, X. Zhang, X. Yu, P. Chou, Q. Wang. "Anomaly detection in online social networks", *Social Networks*, 34(3), pp: 645-654, 2015.
[51] V. Kumar, S. Kumar, A. Kumar Singh. "Outlier detection - A clustering based approach", *IJISME*, I(7), pp. 383-387, 2013.
[52] G. Cui, H. Gao. "Rough Set Processing Outliers in Cluster Analysis", in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 111-115, 2019.
[53] X. Zhao, J. Liang, F. Cao. "A simple and effective outlier detection algorithm for categorical data", *International Journal of Machine Learning and Cybernetics*, 5 (3), pp. 469-477, 2014.
[54] Y.Y. Yao, "Information granulation and rough set approximation", *International Journal of Intelligent Systems*, 16(1), pp. 87-104, 2001
[55] Y. Qian, J. Liang, Y. Yao, C. Dang. "MGRS: A multi-granulation rough set", *Information Sciences*,180(6), pp. 949-970, 2010
[56] [56] Y. Zhou, J.T. Yao. "A Web-Based Learning Support System for Rough Sets", in *International Conference on Rough Sets and Knowledge Technology*. *Springer, Cham*, pp. 161-172, 2014.
[57] Y. Qian, H. Zhang, Y. Sang, J. Liang. "Multi granulation decision-theoretic rough sets", *International Journal of Approximate Reasoning*, 55(1), pp. 225-237, 2014.
[58] T. Zhang, R. Ramakrishnan, M. Livny. "BIRCH: an efficient data clustering method for very large databases", in *ACM Sigmod Record*, 25(2), pp. 103-114, 1996.
[59] H. Wang, W.X. Zhang. "Relationships between concept lattice and rough set", in *International Conference on Artificial Intelligence and Soft Computing*. *Springer, Berlin, Heidelberg*, pp. 538-547, 2006.
[60] Z.A. Bakar, R. Mohemad, A. Ahmad, M.M. Deris. "A comparative study for outlier detection techniques in data mining." in *2006 IEEE conference on cybernetics and intelligent systems*, pp. 1-6, 2006.

Author Biographies

Sangeetha T has completed M.Tech in Computer Science and Engineering from Dr. M.G.R University and B.Tech from Anna University, Tamil Nadu, India. She is working for VIT University as Teaching cum Research Associate. Her research interests include data mining, database management system and applied soft computing.

Geetha Mary A received her Ph.D. from VIT University, Vellore, India. She has completed M.Tech in Computer Science and Engineering from VIT University and B.E. from University of Madras, Tamil Nadu, India. She is working for VIT University as Associate Professor. She was awarded Merit Scholarship for her best academic performance for the year 2007-2008 during her M.Tech study. She has authored more than 20 journal and conference papers. She has authored book chapters in the area of data mining and artificial intelligence. Her research interests include security for data mining, databases and intelligent systems. She works to empower health care management using computer science. Dr. Geetha Mary is associated with many professional bodies like IACSIT, CSTA and IAENG.