OPTIMAL FEATURE BASED DENSITY CLUSTERING FOR OUTLIER DETECTION IN MULTIVARIATE DATA

Karthikeyan G
Ph. D. Research Scholar, Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India
Sr. System Analyst, UST Global, Chennai, Tamil Nadu, India

Balasubramanie P
Professor, Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India

ABSTRACT

Efficient outlier detection in a large-sized big data environment incurs much of complexity in processing the information and to handle it in a proficient way. For segregating outliers from those normal data items, many of the prevailing methodologies experiences complexity in accordance with the features involved in every single attribute. On recognizing appropriate features associated the characteristics of a data gets defined. The necessity of analyzing all sort of feature escalates the processing time along with memory consumption. As a way out of all of these issues, this paper proposes Optimal Feature based Outlier Factor Model (OF-OFM), an effectual outlier detection approach accompanied with prior feature optimization strategy. Initially, preprocessing stage formats all data instances available in the dataset utilized and deployed in a SPARK architecture. Furthermore, an Ant Colony Optimization gets employed in determining for an optimal set of features among the wholesome feature set available. Generalized Sequence Pattern methodology gets employed for formulating tightly coupled sequential patterns that exclude outliers on the basis of a feature set. Moreover, a density based clustering approach involves in clustering those sequentially associated patterns as a means of forming densely associated clusters. As a final point, Local Outlier Factor based outlier detection methodology involves in discriminating outliers completely from that information processed so far. The efficacy of OF-OFM regarding outlier detection gets exemplified by evaluating Area Under Curve (AUC), CPU utilization time, execution time, detection accuracy and memory consumption against existing outlier detection methodologies. OF-OFM evidently proves to be efficacious than other approaches.

Key words: Ant Colony Optimization, General Sequential pattern, Feature optimization, Local Outlier Factor, Outlier Detection.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=9

1. INTRODUCTION

A data mining procedure involves in uncovering some sort of completely useful information or knowledge from those archived within a database via implication of a non-trivial information abstracting procedure [1] [2]. Those abstracted set of knowledge tends to become faulty at some instance when there exists a chance of including some ambiguous or any other deceptive information. Hence, all those information are meant to be alleviated owing to production of adverse consequence in the entire process of either business or a research oriented decisions [3] [4]. Recognizing a pattern that does not comply with usual patterns prevailing in any of the disciplines gets confined as outliers. The cause inferred for generation of outliers that typically gets deviated from those normal set of entries are given as inappropriate sensors, erroneous information entries, typographical error and noise. Their existence completely alters the customized behavior of the system that is found predefined. An efficacious practice of alleviating outliers is highly necessitated in varied fields of medical opinion oriented biological diagnosis as well as other fields that demand a fraudulent identification strategy. A diverse set of approaches prevalently available for identifying outliers is stipulated on the basis of several categories sorted as unsupervised, semi-supervised and supervised. At some instance both sort of supervised as well as semi-supervised methodologies [5] [6]. Outlier recognition by means of utilizing statistical outlier recognizing approaches proceeds with an accomplishment of a priori model constructed on the basis of a probabilistic approach. Hence, the model concerned is formulated in accordance with varied distribution law that exploits empirical information for articulating probability distributions that certainly induces an added complexity in it owing to priori knowledge requirement. Utilizing regression analysis [7] [8] for processing and segregating outliers from those normal entries is achieved via estimating conditional probability distribution. Probably the results with an ultimate accuracy are accomplished only by means of appropriating a robust distribution assessed on the basis of density functionality and it goes for for a non-trivial functionality.

Utilizing kernel functionality based methodology for recognizing outliers employs a feature space for realizing non-linear mapping of objects residing within the input space [9]. The procedure of mapping is accomplished without any sort of external computation instead it is implied in an implicit manner. Hence, the necessity of storage space to opt in addition is completely alleviated. The decision functionality utilized in assessing the outlier performed in a poor manner. It is found unable to segregate the outlier residing within a normal data item entry owing to an exploitation of the decision function to employ outlier segregation procedure. Since it is unfeasible to get deployed on a huge-sized data processing environment or any big data oriented applications in accordance with the complexity inferred in making decisions. Outliers are also proficiently recognized and segregated by means of utilizing fuzzy based kernel functionality technique. This approach deploys a fuzzy set theory in order to resolve a varied set of multi-class classification issues. Both fuzzy SVM as well as fuzzy clustering methodology [10] is utilized in resolving outliers with an enhanced efficacy. The complex portion involved in executing this approach for alleviating outliers is the fabrication of kernel matrix functionality involved getting operated with the learning set inferred. Hence, the features involved with those data instances must be analyzed and optimized in order to mitigate the overall computational complexity [11].
The limitations inferred on the whole are given as,

- Up surged deficiency in detecting outliers in accordance with varying features involved
- Raise of computational complexity with respect to the increase in the data dimensionality of the information involved in a big data environment.
- Enlarged consumption of physical utilities for accomplishing a robust processing of data instances.

In order to overcome these restrictions an inventive feature selection based Optimal Feature based Outlier Factor Model (OF-OFM) methodology is devised.

- An optimal feature selection methodology is endorsed via Ant Colony Optimization methodology on a preprocessed dataset.
- Those optimized best features are further subjected to form a sequential pattern and hence, the tightly associated patterns are formulated and the chance of an outlier occurrence is highly mitigated
- Furthermore, a density based clustering strategy is employed on those patterns in order to alleviate the clusters and a related patterns are clustered
- Finally, the LOF based outlier detection methodology is employed to detect outliers by means of detecting the local reachability distance both in and around clusters and hence the outlier in one cluster may get relevant in another. Thus outliers are completely segregated in a robust manner.

This paper is organized as follows: Section II describes the related works on outlier detection. Section III discusses the proposed Optimal Feature based Outlier Factor Model (OF-OFM). Section IV presents the performance analysis of OF-OFM regarding the detection accuracy, execution time, CPU utilization time, memory consumption and Area Under Curve (AUC) over the existing methods. Finally, section V presents the conclusion.

2. RELATED WORK

This section describes the influence of features involved in the data instances and challenges realized in accordance with the complexity in processing that information in detecting outliers. Filzmoser [12] defined a Mahalanobis distance based outlier detection approach for analyzing multivariate information found from a normally dispensed. It was also capable of processing the data obtained from those asymmetric distributions too. Perozzi, et al. [13] presented an inventive mechanism for accumulating preferences stipulated by a user in order to segregate outliers and to perform clustering on attributed graphs by means of utilizing FocussCO algorithm for processing exemplar nodes offered by users. Though the methodology defined was found highly scalable, and was capable of abstracting focused clusters in a proficient manner, it was lagging in segregating outliers through heterogeneous set of information. Dang, et al.[14] formulated a mathematically deployed model for unveiling outliers from a dimensionally mitigated subspace environment accompanied with localized structure preservation mechanism. Though it was capable of abstracting discriminative set of features in order to segregate outliers, it required a prior processing of information with a labeling procedure. Liu, et al.[15] put forward a proficient methodology for recognizing and segregating outliers found within uncertain information by means of utilizing an innovative Support Vector Data Description (SVDD) mechanism. The outliers were unveiled in a two-fold manner given as articulation of pseudo training step trailed by confidence score allocation. Though it was proficient in abstracting outliers and it was not scalable to process the information residing in a huge flow of information. Radovanović, et al.
[6] endorsed an unsupervised category based reverse nearest neighbor strategy for effectual exposure of outlier in a high-dimensional data forum. Fabricated AntiHub methodology for segregating outliers from data item possessed complete normal behavior. Though the process of distinguishing outliers scored an enhanced accuracy, the execution time acquired for doing so was too long and if mitigated an unhealthy tradeoff was induced between the detection accuracy as well as the execution time.

Upasani and Om [3] devised a fuzzy min-max approach based outlier identification that completely relied upon the principle of majority voting. A threshold preferred by a user was utilized to process the information in the testing stage along with the production of fuzzy membership for every distinct pattern that ensued for an enhanced detection accuracy. However, the limitation was observed with a prolonged inference of recall time in accordance with the time complexity necessitated for voting computation. Zimek, et al.[16] suggested a robust segregating approach for outlier excavation by means of employing an ensemble based methodology and diversity gain assessment. Though it is efficacious in defining outliers, it incurred a large computational complexity without attribute normalization and it was unsuitable for getting implied in diverse set of large-scale applications. Bhuyan, et al.[17] exemplified a tree-based clustering methodology along with a feature abstraction methodology in a fast distributed approach accumulated in a multi-step approach for segregating outliers. Features were abstracted by means of utilizing highly associated non-redundant approach. Though it was capable of probing for outliers in a standard set of unvaried information it was unfeasible to segregate outliers from a categorical kind of information that was reformed periodically.

Bhuyan, et al. [18] surveyed and projected a completely inclusive structure for detecting outliers found in the intrusion dataset as a means of preventing intrusions. Though it was capable of alleviating outliers in a proficient manner, it is incapable of working within a heterogeneous environment. Kontaki, et al.[19] evaluated distance-based approach for detecting outliers involved in a stream based environment. It was capable of fulfilling the storage overhead along with a robust flexibility and effectual processing irrespective of the input parameters involved. The limitation inferred in this approach was its inability to deal with uncertain data in accordance with the surging attributes inferred. Leys, et al.[20] analyzed and projected a dispersion assessing methodology by means of employing median absolute deviation strategy. Though the endorsed approach was completely sufficing the need with a mitigated complexity in terms of implementing, the accuracy involved in predicting outliers was completely relying upon the prior decision generated in accordance with the data instances.

Ienco, et al.[21] presented an inventive mechanism for segregating outliers using semi-supervised outlier recognition approach. Employed a distance based methodology for categorizing normal data instances from those abnormal ones through utilization of a discriminative prototype. Though it was capable of alleviating outliers with a proficiency the time incurred for identifying and distinguishing those outliers was too large. Chen, et al.[22] suggested two diverse set of methodologies constructed on the basis of Support Vector Data Description (SVDD) typically given as NR-SVDD as well as R-SVDD approach. Though the NR-SVDD approach performed better in terms of recognizing outliers with an enhanced accuracy without any sort of falsified identifications, the procedure for analyzing the characteristic feature was completely lagging and hence, it probably lead to a counterfeit recognitions in case of unknown information. Vembandasamy and Karthikeyan [23] endorsed a robust classification methodology influenced through an outlier detection approach in order to discriminate patients infected with diabetes from those normal ones. It exploited a fuzzy c-
means clustering approach for formulating clusters devoid of including outliers within it. Liu, et al. [24] devised an effectual approach for analyzing the localized behavior of data instance involved in processing outliers given through utilization of LOF computation supplemented with an association of K-means clustering approach. Though the outliers were identified in accordance with the global attributes, feature analyzing capability was highly restricted. Lv [25] developed a replacement as well as outlier recognition approach by means of analyzing the neighborhood correlation between data items by deploying Autoregressive and Moving Average Model (ARMA). Though it was capable of computing trends involved a proficient manner, the decision involved in discriminating outliers introduced an uneven tradeoff with accuracy.

3. OPTIMAL FEATURE BASED OUTLIER FACTOR MODEL (OF-OFM)

This section elucidates various methodologies employed in a progressive manner for segregating outliers from the dataset being given in. Fig. 1 exemplifies the overall flow of the proposed outlier detection methodology. Initially, the dataset given is preprocessed to alleviate the unfilled entries and irrelevant noisy elements existing within the dataset. The preprocessed dataset is further ensued by getting subjected to a feature selection procedure that exploits ACO approach.

![Figure 1 Overall flow of method.](image)

This approach opts for best features that assist in a perfect computation of outliers from those normal entries. The features chosen by ACO mechanism finds its applicability in an appropriate manner and hence, only necessitated features are prevailing within processing information is ensured. After omitting those unfeasible features from the dataset, the information that comprises a set of features alone is further progressed as an optimal subset
that comprised only of best features alone. Those best set of features is employed to generate a generalized sequence pattern that is capable of abstracting all sorts of frequent data items alone eliminating those singular items found without any association between frequent items. After obtaining those frequent data items, clustering that information in datasets on the basis of density is accomplished. The clusters formulated with these datasets are assessed for its neighborhood density and then the clusters are fabricated and also further extended to articulate a cluster with densely related items. Finally, the outliers are identified by means of utilizing Local Outlier Factor identification methodology. The density metric of each and every cluster is assessed with degree of outlier-ness in order to find out the inappropriately associated data items and to segregate it as outlier.

<table>
<thead>
<tr>
<th>Table 1 Symbols and Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symbols</strong></td>
</tr>
<tr>
<td>$D$</td>
</tr>
<tr>
<td>$Pd$</td>
</tr>
<tr>
<td>$T$</td>
</tr>
<tr>
<td>$a$</td>
</tr>
<tr>
<td>$C$</td>
</tr>
<tr>
<td>$I$</td>
</tr>
<tr>
<td>$Sf$</td>
</tr>
<tr>
<td>$Sbf$</td>
</tr>
<tr>
<td>$E$</td>
</tr>
<tr>
<td>$Da$</td>
</tr>
<tr>
<td>$Sp$</td>
</tr>
<tr>
<td>$Eps$</td>
</tr>
<tr>
<td>$Mp$</td>
</tr>
<tr>
<td>$Di$</td>
</tr>
<tr>
<td>$Nn$</td>
</tr>
<tr>
<td>$Noc$</td>
</tr>
<tr>
<td>$Ch$</td>
</tr>
<tr>
<td>$N$</td>
</tr>
<tr>
<td>$Ad$</td>
</tr>
<tr>
<td>$Lrd$</td>
</tr>
</tbody>
</table>

3.1. Data Preprocessing

A characteristic feature of cell nuclei is observed on the basis of computing methodology organized through a deployment of an aspiration generated via the finest needle. Those real-valued features are accomplished from a digitized form of an image that is completely utilized for a cytological analysis. The numerical values resembling those characteristic feature of multivariate information prevailing in the dataset is preprocessed to get rid of noisy elements as well as unfilled entries found in the dataset. On preprocessing the overall proficiency of implied data mining procedure gets escalated in a consistent manner. Hence, the characteristic feature of the breast mass is precisely assessed for its malignancy or benign tendency. Thus every distinct imagery residing in the dataset are processed by means of utilizing those input entries and the outliers involved are also alleviated through the trailing procedures.
Preprocessing the Input Dataset
Input: Multivariate Dataset
Output: Preprocessed Dataset

for i=0: D do
    C=0
    for j=0: Ta do
        if a.j == 0 then
            C++
        end if
    end for
    if C == 30 then
        Delete D.i
    end if
end for

Absolute Conversion
for i=0: D do
    for j=0: Ta do
        if a.j < 1 then
            a.j  format(a.j)
        else
            a.j  round(a.j)
        end if
    end for
end for

On handling those data instances prevailing within a dataset to be processed, the data instances are primarily initialized to and are acquired up to a count stipulated in prior for 30 numbers. In particular, in this specified procedure of preprocessing, the numeric values indicating the characteristic feature of the nuclei tissue is completely processed to form a useful measure without any kind of flaw involved in it. At once when the set of attribute values meant for processing reaches 30 a complete data is read and excess values are transformed to next iteration. After consuming those first data instance in the dataset, the process of formatting those values is initiated. In the process of formatting it, those negative values and float values, are formatted to zero or a rounded off value. On manipulating all those numerical values into formatted whole number values, further processing of characteristic features is eased along with an enhanced accuracy. Until entire dataset is processed, the same manner is reiterated for formatting all those attributes residing within the dataset is formatted.

3.2. Feature Selection using ACO
The appropriate set of features are chosen from a dataset that is preprocessed in prior is further processed by utilizing a probability based functionality in order to form an optimal set of features alleviating other irrelevant feature. The probability of formulating an optimal set of items is accomplished by means of computing following calculations given as,

\[ P_b(d) = \left( \frac{\rho}{\rho + E} \right)^y \] (1)

\[ P_b(d) = \begin{cases} 2f(d) & \text{if } E < \rho \\ 1 & \text{if } E \geq \rho \end{cases} \] (2)

Where,
\( \rho \)-similarity coefficient
Optimal Feature Based Density Clustering for Outlier Detection in Multivariate Data

$d$-threshold defined for a data being chosen

$E$ - local density function defined for the chosen data item on the basis of entropy calculated

Both $\rho$ as well as $E$ is found as the base for defining threshold in order to prefer a data. Periodic modification in the value of $\rho$ is acquired for every successive iteration on acquiring some of the productive data by every distinct ant ($I$). The variation of $\rho$ is predicted as,

$$E = -\sum_{k=1}^{n} p_j \log p_j$$

(3)

Where,

$p_j$ – indicates the probability of obtaining a $j$th item from a dataset in an arbitrary manner

The efficacious practice of picking as well as clustering of data items on the basis of features selected by ant is projected in accordance with the $\rho$ specified inside a stipulated area of(0,1). The data items located beyond this specification is left unpicked and is deliberated as failure cases.

Failure in picking data items are given as,

$$\rho^I(t + 1) = \begin{cases} 
 (\rho^I(t) + 0.01), & \text{if } \frac{j^I_j(s)}{j_f} > 0.99 \\
 (\rho^I(t) - 0.01), & \text{if } \frac{j^I_j(s)}{j_f} > 0.99 
\end{cases}$$

(4)

Those clustered items are further processed to obtain neighboring clusters located contiguously by means of computing group average proximity in a basis of the minimal vicinity in location. As a means of accomplishing the knowledge regarding vicinity criterion between data items, any two proximal data items meant to be clustered are computed for average distance and pairwise proximity. Any two neighboring clusters that are positioned at vicinity are recognized by means of exploiting minimum group average proximity condition. Those clusters particularly existing at the vicinity are recommended to the ants who are involved in processing those data items along with its attributes necessitated. The road map that assists ants in search of clusters residing in the vicinity is suggested by the proximity function computed as in (5). The clusters $D_l, D_j$ are those ones who are all positioned in proximal distances and possesses a size measuring $a_w$ and $a_x$ respectively.

$$\text{Proximity} (D_l, D_j) = \frac{\sum_{f} a_w \text{proximity}(f,g)}{a_w \times a_x}$$

(5)

**Selecting best features using ACO**

*Input: Preprocessed Dataset, randomly selected features*

*Output: Best features opted*

for i=0: I do

for j=0: Sf do

$$E = -\sum_{k=1}^{n} p_j \log p_j$$

end for

Sort features according to attributes entropy

for j=0: Sf do

$a_j$=Sbf

end for

end for
The preprocessed dataset is obtained as an input for further processing of data instances on the basis of attributes appropriated along with them. The attributes associated with every data item scores the impact component for opting that particular data item. Fig. 2 shows the optimal features selected after applying ACO on the preprocessed dataset.

Initially, a predefined set of features opted in an arbitrary manner is deliberated for processing the data items. Then the entropy value for every distinct feature opted is assessed and on the basis of $E$ value obtained, ants are recommended to choose the feature of a data item for respective iterations. The entropy value computed accounts for robustness of a feature defined for every single data instance. The path laid down primarily is subjected for reiterating the procedure and it particularly states the difference inferred between the destination and the present feature opted. For such an extent, the features are chosen and entropy is assessed respectively. The entropy value inferred is computed on the basis of probability obtained for those features, which is relying upon the robustness of feature concerned. After assessing the entropy defined, features are sorted in accordance with that associated entropy and finally, the feature with the highest entropy value is picked for each distinct iteration processed. Hence, the dataset that comprises of overall 30 number of attributes along with its associated class name and identity are deliberated. After employing ACO, an optimal set of 10 attributes along with its class and ID is obtained.

3.3. Generalized Sequential Pattern Extraction
Robust features that are obtained through ACO further gets subjected to a generalized sequence pattern algorithm in order to accomplish a sequential pattern that sounds completely associated. As a means of implying such a sequence pattern procedure on those distinct attributes acquired out of previous processing, an only sequential pattern of elemental attributes exists within and those other attributes existing out of sequence are completely alleviated out of processing. This practice certainly eases abstracting procedure of highly associated information organized in a pre-defined manner among those distinct set of features.
availed from the whole dataset. Thus the attributes interlinked are compared with one another on the basis of their characteristic features involved.

**Generalized Sequence Pattern Extraction**

Input: Distinct Attributes  
Output: Selected sequence Pattern

```plaintext
for i=0: Da do  
for j=0: Da do  
  for k=0: A do  
    for l=0: A do  
      if Da.i==A.k && Da.j==A.l then  
        C++  
      end if  
    end for  
  end for  
  if C>0 then  
    Sequencepattern1(Da.i, Da.j)  
  end if  
end for  
end for

Sequencepattern1(Da.i, Da.j)  
for k=0: A do  
for l=0: A do  
  if Da.i==A.k && Da.j==A.l then  
    Sp ← selected pattern  
  end if  
end for```

Those distinct attributes are acquired in an order in which it is presented in the database. Each and every attribute obtained is analyzed and archived in a sequential pattern to be framed. Analysis of those attributes is carried out in a certain way that characteristic feature of the attribute concerned is compared with each other, if it is found to be similar, then it is merged in a selected set of a pattern. Hence, the selected pattern comprises only of completely associated data instances on the basis of its similar behavior.

### 3.4. Density based Frequent Pattern Clustering

The density based clustering procedure is applied to cluster those most similar set of sequence patterns acquired and to preserve it in a contiguous manner. The minimal radial distance in which those correlated clusters found is assessed initially on the basis of sequential patterns being formed. Each and every pattern is assessed for its characteristic features. The distance between those similar clusters is assessed using eps ($D_i$) that define the radial distance between clusters. This $D_i$ measure defines the vicinity between clusters in accordance with its characteristics. If the distance between any two patterns is lower than the $D_i$ metric, then that particular item in the patterns is deliberated as nearest neighbor. Likewise, all other data items found inside the pattern are analyzed and each and every similar pattern are merged into clusters on the basis of nearness inferred in agreement with its characteristic feature. Such appending of similar patterns certainly reaches a threshold limit when the count of minimal set of points necessitated to form a dense region is reached.
Density based Clustering

Input: Selected sequence Pattern
Output: neighboring clusters

for i=0: D do
  for j=i+1: D do
    \[ D_i = \sqrt{(D_i - D_j)^2} \]
    if \( D_i < \text{eps} \) then
      \( N_n \leftarrow D.i \)
    end if
  end for
  if \( N_n > M_p \) then
    expandcluster()
  end if
end for
expandcluster()

for i=0: N_n do
  for j=i+1: N_n do
    \[ D_i = \sqrt{(D_i - D_j)^2} \]
    if \( D_i < \text{eps} \) then
      \( N_n \leftarrow D.i \)
    end if
  end for
  if \( N_n > M_p \) then
    \( N_n \leftarrow \text{merge neighbors} \)
  end if
end for

The maximum limit for which the minimal set of points for framing a dense region is confined on the basic criterion stating that nearness between patterns residing in a cluster is completely associated with characteristics of another clustered pattern. At this juncture, the cluster is to be expanded on the characteristic feature newly inferred. The similar procedure is reiterated until all sorts of associated patterns recognized are tightly clustered on the basis of the density function. Additionally, the neighbors found around those newly expanded are clustered in accordance with its characteristic feature.

3.5. Outlier detection with Local Factor

The outlier factor defined for a cluster is devised on the basis of minimized nearest points found within the nearest reachability density points. The concept dealt here works relying upon the concept given as the closer distance of the minimum point, the higher the local reachability of densities found in between them. Many of the characteristics defined for those local outliers are desired as follows:

- The LOF defined for a cluster is mitigated and estimated almost equivalent to the numerical value 1.
- A lower and upper similarity bound of a cluster is defined for all those objects found beyond the scope of associated clusters
- Ultimate tightness of those bounded clusters is assessed
- The tightness among clusters are very high in accordance with the similarity assessed among clusters with patterns
The tightness of those bounds are loosely coupled if the distance between those objects belonging to other clusters are encountered.

The LOF of an object appropriated relying upon those minimum reachability distance inferred between any two similar objects. Both minimum and maximum reachability is actually defined as,

\[
reach - dist - min = \min \{reach - dist (p, q) \mid p, q \in C \} \quad (6)
\]

\[
reach - dist - max = \left( \frac{reach - dist - max}{reach - dist - min} - 1 \right) \quad (7)
\]

In general, any of the node \( N \in Ch \) is given as,

- All those min-pts nearest to any neighbor \( q \) that belonged to an object’s node \( N \) certainly gets confined to \( Ch \)
- Any of the other min-pts nearest to any neighbor \( o \) of any object \( N \) also resides within the reach of \( Ch \) and is shown as,

For all sorts of node object \( N \) residing inside a \( Ch \) is possessing a local reachability density of that \( Ch \) concerned and is given as,

\[
N \leq \frac{1}{reach - dist - min} \quad (8)
\]

Conversely, the maximum distance inferred for an object to define its reachability is \( reach - dist (p, q) \leq reach - dist - max \),

\[
N \geq \frac{1}{reach - dist - max} \quad (9)
\]

Hence, the LOF for all patterns involved is defined as,

\[
\frac{1}{(1 + \varepsilon)} \leq LOF (p) \leq (1 + \varepsilon) \quad (10)
\]

So, for an object node found inside the cluster head \( Ch \) is always counted for a deep LOF oriented objects for a highly tight cluster almost close to 1.

**Local Outlier Factor Detection**

Input: Clustered patterns
Output: Outliers

```
for i=0: Noc do
    for j=0: N do
        Di = \sqrt{(N_j - Ch_i)^2}
    end for
    Ad=Di/N
end for
for j=0: N do
    Lrd+= 1/(Di/Ad)
end for
```

Fig. 3 shows the outcome obtained after detecting outlier and removing it to obtain only the useful information.
For all sorts of clusters formed using varied set of sequential patterns, every single node is deliberated for processing in an individual manner. The distance between each and every node is assessed average distance is computed. On assessing the distance between each and every node \((N)\) and those cluster heads \((Ch)\) the minimum points are defined and further average distance \((Ad)\) is computed. On calculating the average distance, the between a single object and all other clusters the outlier in a cluster may get converged with another relevant cluster. Hence, all kinds of objects are tightly bounded. The \(Lrd\) is assessed for and is density is discovered for accomplishing a completely associated cluster and hence, the leftover objects are deliberated as outliers. Finally, the normal data instances without any sort of outliers are preserved and outliers are alleviated from it.

4. PERFORMANCE ANALYSIS
The performance validation of proposed OF-OFM over the existing outlier detection techniques is presented in this section.

4.1. Dataset Description
The Wisconsin Diagnostic Breast Cancer (WDBC) [26] considered for result manipulation comprises of several attributes given as ID number, diagnosis of the cell nuclei imagery inferred for analysis of breast mass. All those characteristic features are transformed and represented as a numerical values given as smoothness, compactness, perimeter, radius, concavity, area, symmetry and concave points acquired from a digitized imagery. WDBC is consists of 569 samples of instances appropriated with 32 attributes residing in a multi-variate dataset characterizing cell-nuclei of an image. SPARK architecture is utilized to realize and process this dataset in a big data environment that is capable of managing huge-sized data in a proficient manner.

4.2. Area under the Curve (AUC)
The OF-OFM algorithm examines the performance of analyzing the association inferred between normal patterns along with those abnormal patterns on the basis of accuracy and error predictions [21]. Abnormal occurrences are efficaciously unveiled in accordance with
the detection rate exposed by the proposed algorithms. Detection rate also determines the value of AUC and AUC surges with respect to higher the detection rate inferred. The comparison research study [27] showed that the SANDCat offered the high AUC value compared to the existing One-Class Support Vector Machine (OCSVM), unconstrained Least Square Importance Fitting (ULSIF), Feature Ensemble model (FRaC) and Local Outlier Factor (LOF). The comparative analysis of AUC for various existing methodologies against the devised OF-OFM approach is illustrated in Fig 4.

\[
AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0n_1}
\]

Where, 
- \(S_0\) - The sum of rank of abnormal instances detection 
- \(n_0\) - The number of test instances belongs to normal 
- \(n_1\) - The number of instances belongs to the outliers

Among those prevalent methodologies found, FRaC exposes a better performance and it is certainly outperformed by the devised OF-OFM by 22.09%.

4.3. CPU Utilization Time

The effectual practice of utilizing CPU for processing a given cluster along with its associated features. The effectiveness of the proposed OF-OFM algorithm is completely governed by prior feature mitigation procedure accomplished by the ACO mechanism implied on those preprocessed datasets [28]. Even though the entire information setup is realized in a big data environment, the features are skillfully reduced the dimensionality of the data set and thus time consumed for processing an entire cluster is less. The comparison of the CPU utilization time for the proposed OF-OFM and the existing Back Propagation Neural Network (BPNN), Artificial Neural Network and Fuzzy Clustering, Hyperbolic Hopfiled Neural network, Neighborhood Outlier Factor methods shows that the OF-OFM offers the less CPU utilization time.
For proposed OF-OFM, the CPU utilization time for the minimum dataset size of 5000 and maximum dataset size of 25000 is 6 ms and 15.5 ms respectively. Conversely, the neighborhood outlier factor that outperforms among the prevailing methodology exposes a CPU utilization time of 9 ms and 20 ms for the dataset size of minimum and maximum of 5000 and 25000 respectively. Fig 5 shows that the OF-OFM reduces the CPU utilization time by 33.33 % and 22.5 % for minimum and maximum dataset size.

4.4. Execution Time

The overall time taken for segregating those outliers from the normal data are designated as execution time. The effectiveness of the proposed OF-OFM methodology relies upon the feature reduction that completely overcomes the problem of “Curse of dimensionality” through applying ACO algorithm. Prior feature reduction gradually mitigates the overall execution time for the entire process. Fig. 6 illustrates the execution time for various prevailing methodologies such as Back Propagation Neural Network (BPNN), Artificial Neural Network and Fuzzy Clustering, Hyperbolic Hopfiled Neural network, Neighborhood Outlier Factor against the endorser OF-OFM approach on the basis of varied dataset size.
Among those prevailing methodologies, neighborhood outlier factor approach exposes a proficient performance for the minimum dataset size of 5000 and maximum dataset size of 25000 with an execution time of 40000 ms and 190000 ms respectively. While, the proposed methodology further mitigates the execution time for the same amount of dataset size by giving execution time as 27000 ms and 165000 ms respectively, which is proficient by 32.5 % and 13.15% respectively.

4.5. Detection Accuracy

The proportion defined between true outliers defined to that of the total count of identifications made on the whole specifies the detection accuracy inferred.

\[
\text{Accuracy} \, (\%) = \frac{\text{No of true outlier detections}}{\text{Total number of detections}} \times 100
\]

The accuracy gets surged in accordance with the escalation in true detections made for identifying outliers and the utmost accuracy is exposed by the proposed OF-OFM approach owing to best feature selections made by applying ACO and sequence pattern clustering methodology.

![Figure 7 Detection Accuracy analysis](image)

The detection accuracy of the OF-OFM technique is compared against Memory Efficient increment Local Outlier Factor detection algorithm for more flexible version (MiLOF_F) MiLOF-F mechanism and is made for various data dimensions ranging from 10 to 50 in a specific instances of 10. The accuracy value exposed by MiLOF-F for minimized data dimension is 94 % and for the maximized data dimension is 93%. The proposed OF-OFM efficiently reduces the computational complexity by using the optimal feature selection techniques and provides the accuracy up to 97 % for data dimension of 10 and 99.64% of a data dimension of 50. Fig 7 shows that the effective detection accuracy obtained and the deviation estimation in OF-OFM offers 3.09% and 6.66 % better than the MiLOF-F respectively for minimum and maximum data dimensions.

4.6. Memory Consumption

The perfect operating functionality of a system is accomplished by proficient memory usage in order to complete the processing instances within a stipulated time for all sorts of datasets being employed. Differing datasets that range from 10 to 50 of dimensions is processed by both OF-OFM and MiLOF_F approaches. As depicted in Fig.8 the proposed OF-OFM methodology necessitates only the least amount of memory than the existing MiLOF_F in a comparative manner.
The proficiency of the formulated OF-OFM relies on earlier attribute reduction depending upon the features opted and hence, the entire memory consumption is mitigated on the whole. Prevailing MiLOF_F comparatively utilizes a minimized amount of memory than other approaches for processing and archiving the information. For the minimized data dimension of 10, the amount of memory consumed is 0.1 MB and for the maximum data dimension of 50, the amount of memory consumed is 0.23 MB for MiLOF_F technique. Conversely, the endorsed OF-OFM methodology exposes a memory consumption of 0.03 MB and 0.16 MB for data dimensions 10 and 50 respectively. Ultimately, OF-OFM exposes 56.52% and 81.25% reduction than MILOF_F.

5. CONCLUSIONS
This paper proposed an Optimal Feature based Outlier Factor Model (OF-OFM) for proficient outlier detection approach by means of easing the processing procedure involved in. This methodology is capable of segregating outliers by means of employing an optimal feature optimization methodology through ACO implying on a preprocessed data. The entire dataset is recognized in a SPARK architecture. After a Generalized Sequence Pattern methodology is employed for formulating tightly coupled sequential patterns that are certainly capable of excluding outliers based on feature set. Moreover, a density based clustering approach is involved in clustering those sequentially associated patterns as a means of forming densely associated clusters. To end with, the Local Outlier Factor based outlier detection methodology is employed in discriminating outliers completely from that information being processed. The efficacy of OF-OFM regarding outlier detection gets exemplified by evaluating Area Under Curve (AUC), CPU utilization time, execution time, detection accuracy and memory consumption against existing outlier detection methodologies. OF-OFM evidently proves to be efficacious than other approaches in terms of mitigated execution time by 32.5% and 13.15% for the minimum and maximum dataset size accompanied with an enhanced detection accuracy of 3.09% and 6.66% for minimized and maximized data dimensions.

REFERENCES
Optimal Feature Based Density Clustering for Outlier Detection in Multivariate Data


