OFS-NN: Optimal Features-Neural Network Based Outlier Detection for Big Data Analysis

Karthikeyan G[1, 2] and Balasubramanie P[1]
[1]Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode-638 052, Tamil Nadu, India
[2] UST Global, Chennai-600096, Tamil Nadu, India
Email:pbalu_20032001@yahoo.co.in

***Abstract*—**Outlier detection is massively consist of large number of application domain. Previously the outlier is known as noisy data, but in recent days it became more important in various field because of its usage to detect the unpredicted and unidentified. The outlier detection are used in certain areas like fraud detection of credit cards, calling cards, discovering computer intrusion and criminal behaviors etc. Aim: the main motive of this research is to propose feature selection and subset based outlier detection. Methods: This paper proposes Optimal Feature selection based Neural Network (OFS-NN) 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. The datasets are preprocessed and divided into subsets. Furthermore, an Artificial Bee Colony Optimization gets employed in determining for an optimal set of features among the wholesome feature set available. Artificial Bee Colony that exclude outliers on the basis of a feature set. Moreover, an Expectation maximization clustering approach involves in clustering maximum similar data. As a final point, the neural network classification is used for outlier detection. Results: The efficacy of OFS-NN 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. OFS-NN evidently proves to be efficacious than other approaches in terms of mitigated execution time under minimum and maximum dataset size.

***Index Terms*—**ABC algorithm, expectation maximization clustering, neural network classification

I. Introduction

The concept of outlier is to detect a small set of object which deviated from the common set of object [1]. The challenging issues in outlier detection is the outlier object is small as compared to the regular object [2]. All existing studies mainly focus on the problem in the outlier detection but forget to explain the outlier explanation [3]. In recent times the outlier detection had well grow with the help of hardware and software technologies it enabled with various forms of temporal data collection. The data mining using classification used for the prediction students' performance. The main aim of work is to use the data mining to study the student's performance. The classification task is used to evaluate the performance. The decision tree method is used to get effective results [4]. The huge data availability in wireless sensor network leads to problem of extracting knowledge from the spatio-temporal data. This work mainly deals with outlier detection in spatio-temporal data. The proposed method is Rough outlier set extraction based on the rough set. The fuzzy clustering algorithm which obtained better results with less computational time [5]. The outlier detection is most important research and application domain. But in the real time application the uncertain in nature of data leads to various problems. The proposed method is SVDD –based approach to detect outlier in uncertain data [6]. The clustering guided inadequate auxiliary learning (CGSSL) is proposed by coordinating group examination and meager basic investigation into a joint structure and trial. The bunch names are additionally anticipated by abusing to inadequate model [7]. Counterfeit Bee Colony (ABC) calculation is another swarm knowledge strategy which reenacts insightful scavenging conduct of bumble bees. In the ABC algorithm, there are three types of honey bees namely utilized honey bee, spectator honey bees and scout honey bees. The main aim of the bees to improve the quality of the honey and to get best source of nectar as like the same the algorithm give the best performance and quality results [8].

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 selection based Neural Network outlier detection (OFS-NN) methodology is devised.

- The input dataset is preprocessed and that datasets are divided into subsets.
- The feature selection take place in the subset by using the Artificial Bee Colony algorithm.
- After the feature selection, the datasets are cluster into groups by using the Expectation maximization clustering.

Manuscript received January 2, 2018; revised June 20, 2018.
Corresponding Author:kathiag@gmail.com
doi:10.12720/jcm.13.7.396-405

©2018 Journal of Communications

- In the cluster datasets the outlier detection and removal are made by the neural network classification.

This paper is planned as follows: Section II defines the related works on outlier detection. Section III discusses the proposed Optimal Feature- Neural network based on Outlier detection (OFS-NN). Section IV presents the performance analysis of OFS-NN 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.

II. Related Work

Radovanovic et al [9] Proposed a Reverse Nearest - Neighbors in Unsupervised Distance-Based Outlier Detection. Turn around nearest neighbors in the unsupervised exemption distinguishing proof setting. This strategy accomplished high dimensionality however the precision is diminished when the execution speed is expanded. *Perozzi et al* [10] Introduced the Focused Clustering in Outlier Detection for Large Attributes .The key part of our approach is to construe client inclination by the purported center traits through an arrangement of client gave model hubs. It demonstrated the adequacy and adaptability of the proposed technique on engineered and genuine diagrams yet not reasonable to deal with the recently postured issue. *Hubert et al* [11] Suggested a multivariate practical information examination by new systems for exception identification. The instruments incorporate factual profundity capacities and separation measures gotten from them. The strategies are relative invariant in p - dimensional space, and don't expect circular or whatever other symmetry. The instrument can't be connected for directed arrangement of multivariate information and of useful information. *Liu et al* [12] Proposed the presents a novel anomaly location way to deal with address information with flawed names and join constrained anomalous cases into learning. The proposed methodologies can accomplish a superior exchange off between identification rate and false caution rate. *Fleischhacker et al* [13] Introduced an approach which joins the results of two free anomaly discovery hurries to get a more solid outcome and to likewise keep issues emerging from regular anomalies which are extraordinary esteems in the dataset however all things considered right. The likelihood of proficiently discovering sets of qualities on which multi-variation anomaly location is troublesome. *Zimek et al* [14] Proposed subsampling for effective outlier detection. A group of a few exception identifiers on subsamples of the dataset is more proficient than troupes in light of different methods for presenting assorted qualities and, contingent upon the specimen rate and the extent of the outfit. *Dang et al* [15] proposed a calculation that reveals anomalies in subspaces of diminished dimensionality in which they are very much separated from general articles while in the meantime holding the normal neighborhood structure of the first information to guarantee the nature of exception clarification *Liu et.al* [16] Feature choice is a successful strategy to manage high-dimensional information. It proposed an effective semi-managed highlight determination calculation to choose pertinent components utilizing both named and un-marked information. It is fascinating that the worldwide ideal arrangement of the element determination with the clamor uncaring follow proportion paradigm can be a great deal more proficiently obtained. *Lin et al* [17] proposed a novel component portrayal approach, to be specific the cluster focus and closest neighbor (CANN). Approach. It gave a higher precision and identification rates and a lower false caution rate. The benefit of CANN is that it requires less computational exertion than the k-NN or SVM classifiers prepared and tried by the two unique datasets. *Cai et al* [18] proposed a self –sorting approach for spatial anomaly identification. Proficiency and heartiness are the benefits of proposed system. Ferrari *et al* [19] Proposed. Meta-learning brings the idea of finding out about realizing; that is, the meta-information acquired from the calculation learning process permits the change of the calculation execution. The outcomes demonstrate that the new meta-learning gives top notch calculation choice to grouping tasks. *Capozzoli et al* [20] proposed an approach utilizes measurable example acknowledgment procedures and fake neural ensembling systems combined with exception discovery techniques for blame identification. The outcomes demonstrate the value of this information investigation approach in programmed blame discovery by decreasing the quantity of false anomalies. Siddiqui *et al* [21] introduced the Clustering is an intense information digging procedure for information examination specifically when the information are huge and complex. This compelling strategy to investigations and concentrate the example of assault sort. The examination uncovered many intriguing outcomes about the conventions and assault sorts favored by the programmers for encroaching the networks. Cao et.al [22] proposed a classes of novel direction exception definitions that model the bizarre conduct of moving items for a huge scope of continuous applications. *B.van Stein et al* [23] Proposed a novel neighbor-based direction exception definitions. The improved MEX methodology versatile to huge information direction streams to distinguish the new classes of exceptions, rendering moving item anomalies identification pragmatic progressively applications. To empower precise, neighborhood subspace anomaly location in high-dimensional information that is a blend of segment. Along these lines present GLOSS, a calculation that performs nearby subspace anomaly location utilizing worldwide neighborhoods. GLOSS builds upon both Loop and Hics by integrating both algorithms and adapting them to the global neighborhood. *Ienco et.al* [24] Proposed a semi-supervised anomaly detection for categorical data.

The main aim of the work to differentiate the normal and anomalous occurrence.

III. ABC BASED FEATURE SUBSET SELECTION AND OUTLIER DETECTION

This section explains about the overall work flow of the system as shown in Fig. 1. This system consists of preprocessing, subset separation, feature selection, expectation maximization clustering, and neural network classification. The datasets are preprocessed, in which the noise are reduced. The noise in the datasets are completely removed before it separated into subsets. The preprocessed datasets are separated into subsets. The subsets are classified based on their size. The feature selection process take place in the separated datasets using ABC algorithm. The expectation maximization clustering employed to cluster the feature. The clustering is made for grouping the similar datasets. The similar datasets are cluster in groups after the feature selection.

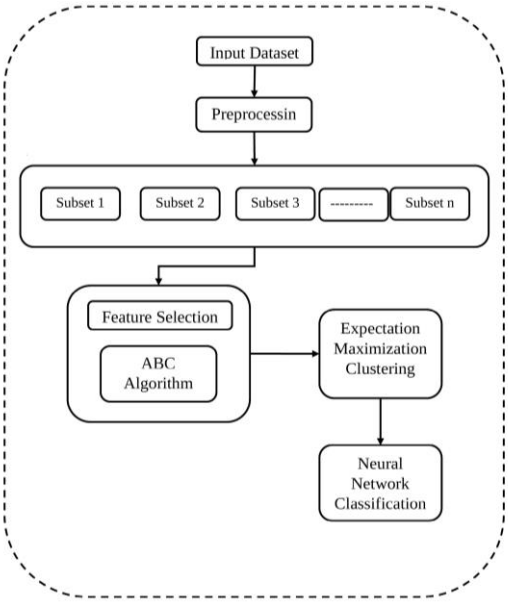


Fig. 1. Overall flow of method.

TABLE I: SYMBOLS AND DESCRIPTIONS

Symbols	Description
D	Load Dataset
Pd	Preprocessed Dataset
T	Total Attributes
a	attribute
C	Count
I	Iteration
Sf	Arbitrarily selected features
Di	Distance
Noc	Number of Clusters
Ch	Cluster Head
N	Nodes
Sc	Selected clusters
Ed	Existing system
A	Average
Ta	*total attributes*
Tv	*Total values*
Hp	*head probability*
Ht	*Head total*
Nd1	*node data 1*
Nd2	*node data 2*
Nd3	*Node data 3*
W	*Weight*
Td	*Total data*
Nos	*Number of subset*
Ss	*Subset size*
Ts	*Total subset*
T	*temp*

A. Data Preprocessing

The real datasets are generally incomplete, noisy and inconsistent. To remove those all the data preprocessing take place. The missing attribute values are filled during the data preprocessing. During the data preprocessing the data cleaning process take place which remove the noisy data, identify or remove outlier. 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

```
for i=0: D do
   C=0
   for j=0: Ta do
            if Ta.j == 0 then
         C++
      end if
   end for
   if C == R then    //R= attributes of the data
      Delete D.i
   else
      Pd ← D.i
            end if
end for
```

Initially the dataset (D) is loaded as input to the preprocessing algorithm. For each attribute i the values are assigned as zero for loaded dataset D and the count value is set as zero. Then for the total attribute Ta the attribute value j set as zero. If the total attribute value is equal to 0 then the count value is increased. When the dataset value is equal to total number of attributes of the data, then the loaded dataset is deleted. Otherwise the dataset is processed as preprocessing dataset.

B. Subset Separation

The preprocessed datasets are split into different subsets. The subsets are spilt in a sequence to perform faster. The subset separation is necessary to decrease the computational time and also help in the feature selection.

Initially the subset size is calculated by dividing the total data with the number of subset. For each value of the subset, the comparison with number of subset (Nos-1) minus 1 is equal to the value then the Ts is calculated. The total value is calculated by total data minus the temp value. Otherwise the subset size is determined with increment of temp value. Then the subset size is maintain as total subset value.

Subset Separation

Ss=Td / Nos
for *i*=0: *Nos* **do**
  if *i == (Nos-1)* **then**
    Ts.i ← *Td-T*
  else
    T += Ss
    Ts.i ← *Ss*
  end if
end for

C. Feature Selection Using ABC

Feature selection is an important pre-requisite for classification. The feature selection process is define as extracting useful information from datasets without any noise data. The procedure of highlight determination requires two imperative parts: Evaluation capacity to assess the hopeful component subset; Generation methodology to create the applicant include subsets.

Artificial bee colony algorithm (ABC) is a popular based stochastic improvement proposed by Karaboga understand the astute scavenging conduct of bumble bee swarms. It can be utilized for characterization, grouping and improvement thinks about. Artificial bee colony algorithm (ABC) comprises of three unique gatherings: utilized honey bees, spectator honey bees, and scout honey bees. In this calculation, the quantity of honey bees utilized in the state likewise rises to the quantity of passer by honey bees. Also, the quantity of utilized honey bees or spectator honey bee's measures up to the quantity of arrangements in the populace. A spectator honey bee is the honey bee that holds up in the move range to settle on the nourishment source determination choice. A spectator honey bee is named utilized honey bee once it goes to a nourishment source. A utilized honey bee that has expended the nourishment source transforms into a scout honey bee, and its obligation is to play out an irregular pursuit to find new assets. A bystander bumble bee evaluates the nectar information of each and every used bumble bee and picks a sustenance source dependent upon the probability of the nectar entirety. Much the same as the utilized honey bees, the passer by honey bee adjusts the circumstance from memory and it checks the nectar measure of the hopeful source. As like the same, the manufactured honey bee settlement calculation work.

Selecting best features using ABC

Ta ← *total attributes*
Tv ← *total values*
A ← *average*
Sf ← *selected features*
for *i*=0: *Nos* **do**

$$Tv = \sum_{j=1}^{Ta} \sum_{k=1}^{Td} Td.j.k$$

$$A = Tv/Ta$$

for *j*=0: *Ta* **do**
  if *Ta.j* > *A* **then**
    Sf ← *Ta.j*
    end if
  end for
end for

The preprocessed subset 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. After the subset separation, the feature selection using the ABC algorithm take place. The total value Tv is calculated using the equation (1) given below:

$$Tv = \sum_{j=1}^{Ta} \sum_{k=1}^{Td} Td.j.k \quad (1)$$

For each subset the total value is calculated. The average value is calculated by using the equation (2)1`2

$$A = Tv/Ta \quad (2)$$

The average is calculated by dividing the total value by total attributes. If Ta is greater than the average A then the total attribute is maintained as selected features.

Fig. 2 shows the optimal features selected after applying ABC on the preprocessed dataset.

featurename	value
M_Radius	14.1272917398946
M_Texture	19.2896485061512
M_Perimeter	91.9690333919157
M_Area	654.889103690686
SE_Perimeter	2.86605922671353
SE_Area	40.337079086116
W_Radius	16.2691898066784
W_Texture	25.677223198594
W_Perimeter	107.261212653779
W_Area	880.583128295254

Fig. 2. Features optimized using ABC

D. Expectation Maximization Clustering

The expectation maximization algorithm is mainly used in estimation problem involving incomplete data. Due to its simplicity, the EM is used in various application such as cluster analysis and statistical pattern recognition. Especially in the outlier detection it act as a tool for detecting abnormal data sets.

Exceptation Maximization Clustering

for *i*=0: *Noc* **do**
  Ch ← *select cluster heads*
end for
T ← *total*
for *i*=0: *Ta* **do**
  $T += \sum_{j=1}^{N} D.i.j$
end for
for *i*=0: *Ch* **do**
  $Ht.i = \sum_{j=1}^{N} D.i.j$
  $Hp.i = Ht.i/T$
end for
for *i*=0: *Td* **do**
  $Nt.i = \sum_{j=1}^{N} D.i.j$
  $Np.i = Nt.i/T$
  for *j*=0: *Hp* **do**
    if *Hp.j – Np.i < Ed* **then**
      Ed= Hp.j – Np.i
      Sc=Ch.j
      end if
    end for
end for

The expectation maximization clustering algorithm helpful in clustering the datasets after the feature selection. After the feature selection the data sets are cluster into separate groups. The Noc is define as number of clusters and ch as the cluster heads. Initially number of clusters and cluster heads are defined for each attributes. For each value the total is calculated by using the equation (3)

$$T += \sum_{j=1}^{N} D.i.j \tag{3}$$

The head total and head probability are calculated by using the equation (4) and (5)

$$Ht.i = \sum_{j=1}^{N} D.i.j \tag{4}$$

$$Hp.i = Ht.i/T \tag{5}$$

The node total and node probability are calculated by using the equation

$$Nt.i = \sum_{j=1}^{N} D.i.j \tag{6}$$

$$Np.i = Nt.i/T \tag{7}$$

If the nodal probability minus head probability less than the existing distance then existing distance is determined as head probability minus nodal probability. Then the selected cluster heads remain as it is. The Fig. 3 shows the exception maximization clustering in feature selection and subset based outlier detection.

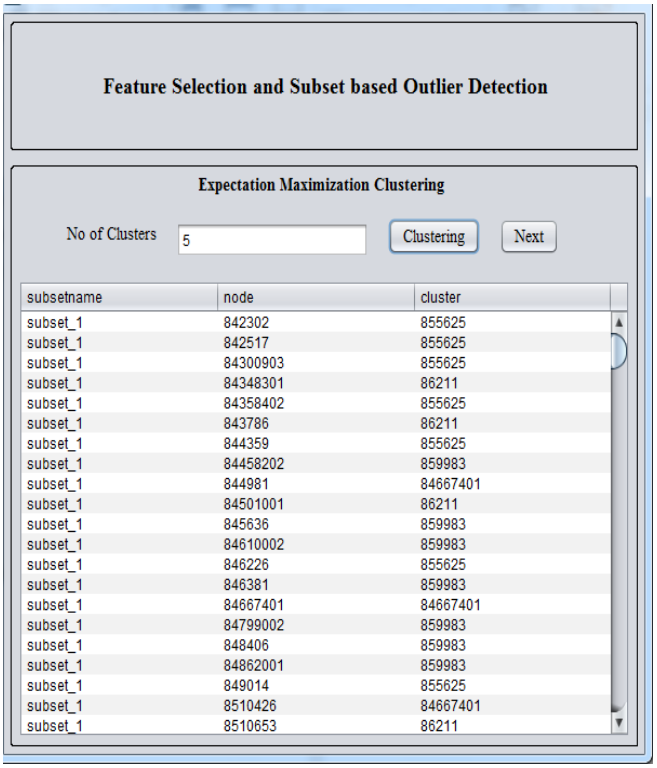


Fig. 3 Expectation and maximization clustering.

E. Outlier Detection Using Neural Network Classification Possessing

Classification is one of the most active research and application area of neural networks. The classification is an important feature to separate huge datasets into classes. Neural Networks are measured as a strong classifier. The neural network classification used for the outlier detection.

Outlier detection using neural network classification

Nd1 ← *node data 1*
Nd2 ← *node data 2*
Nd3 ← *node data 3*
W ← *weight*
for *i*=0: *Noc* **do**
  for *j*=0: *N* **do**

$Nd1 += N.j$

$Nd2 += N.j + W$

end for

Nd3=(Nd1+Nd2)/2

if *Nd1 < Nd3* **then**

Normal

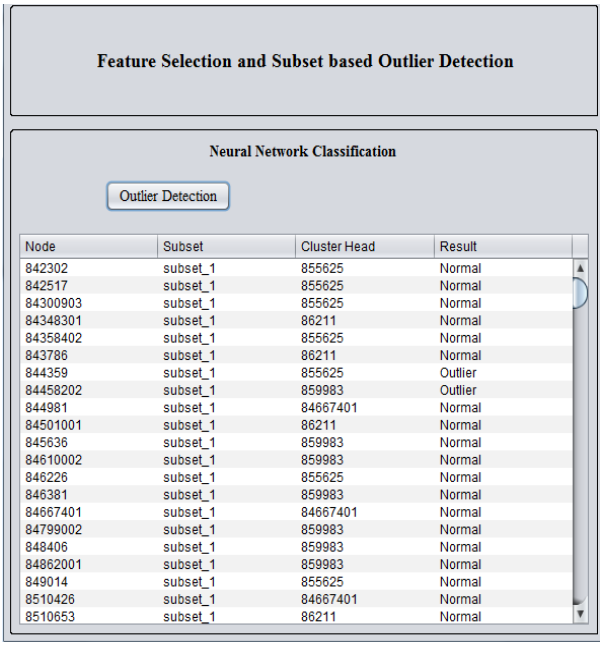
else

Outlier

end if

end for

After the feature selection the outlier detection take place. Initially the data set are separate into node data 1, node data 2, and nose data 3. The first data is calculated as Nd1 as N.j, then the node 2 data is calculated as the first data Nd1 along the weight (W). The weight (W) as the distance between the first node data 1 and node data 2. Then the node data 3 is calculated as the average between the node data 1 and node data 2. If the node data 1 is less than node data 3 then it is normal. Otherwise it considered as outlier. The outlier is identified by neural network classification algorithm. Fig. 4 shows the outcome obtained after detecting outlier and removing it to obtain only the useful information.



Node	Subset	Cluster Head	Result
842302	subset_1	855625	Normal
842517	subset_1	855625	Normal
84300903	subset_1	855625	Normal
84348301	subset_1	86211	Normal
84358402	subset_1	855625	Normal
843786	subset_1	86211	Normal
844359	subset_1	855625	Outlier
84458202	subset_1	859983	Outlier
844981	subset_1	84667401	Normal
84501001	subset_1	86211	Normal
845636	subset_1	859983	Normal
84610002	subset_1	859983	Normal
846226	subset_1	855625	Normal
846381	subset_1	859983	Normal
84667401	subset_1	84667401	Normal
84799002	subset_1	859983	Normal
848406	subset_1	859983	Normal
84862001	subset_1	859983	Normal
849014	subset_1	855625	Normal
8510426	subset_1	84667401	Normal
8510653	subset_1	86211	Normal

Fig. 4. Outlier detection using neural network classification

The main aim of this research is to propose feature selection and subset based outlier detection in the big data environments. The SPARK architecture helps to process the data set in a big data environment. As related with this, the ABC algorithm derived from the characteristics nature of honey bee, helps to select the best features. In addition to that the neural network are measured as a strong classifier, it helps to identify the outlier.

IV. Performance Analysis

The performance validation of proposed OFS-NN over the existing outlier detection techniques is presented in this section.

A. Dataset Description

The Wisconsin Diagnostic Breast Cancer (WDBC) [25] 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, and 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.

B. Area under the Curve (AUC)

The OFS-NN 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 [24]. 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 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 OFS-NN approach is illustrated in Fig. 5.

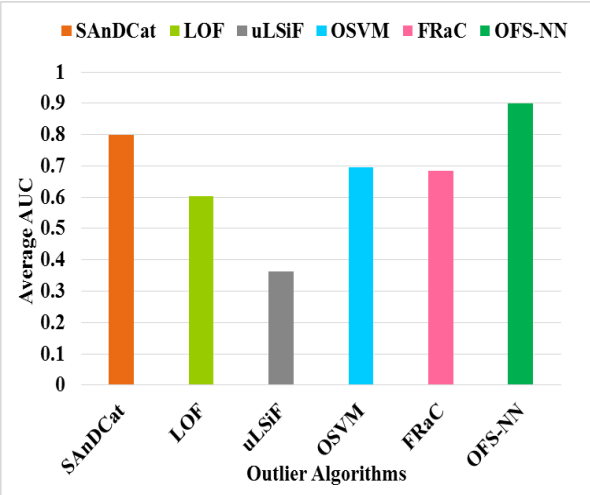


Fig. 5. AUC analysis

AUC is escalated in OFS-NN owing to the variability measure involved in and is mathematically defined as,

$$AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \quad (8)$$

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 OFS-NN by 23.82%.

C. CPU Utilization Time

The effectual practice of utilizing CPU for processing a given cluster along with its associated features. The effectiveness of the proposed OFS-NN algorithm is completely governed by prior feature mitigation procedure accomplished by the ABC mechanism implied on those preprocessed datasets [26]. 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 OFS-NN and the existing Back Propagation Neural Network (BPNN), Artificial Neural Network and Fuzzy Clustering, Hyperbolic Hopfield Neural network, Neighborhood Outlier Factor methods shows that the OFS-NN offers the less CPU utilization time.

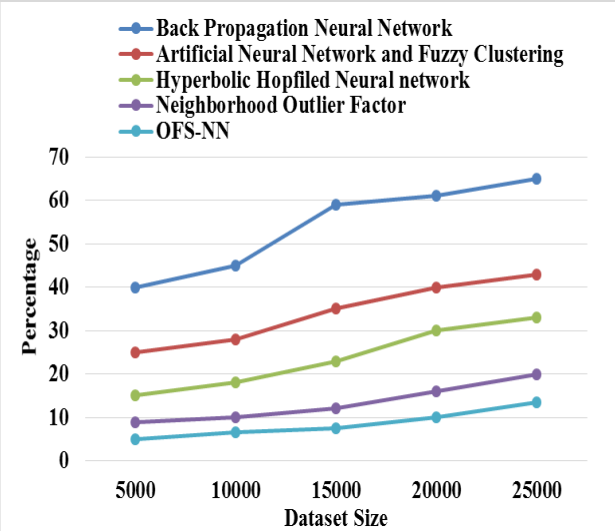


Fig. 6. CPU Utilization time analysis

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. For proposed OFS-NN, the CPU utilization time for the minimum dataset size (5000) and maximum dataset size (25000) is 5 ms and 13.5 ms respectively. Fig. 6 Comparative analysis shows that the proposed OFS-NN reduces the CPU utilization time by 44.44% and 32.5 % for minimum and maximum dataset size.

D. Execution Time

The overall time taken for segregating those outliers from the normal data are designated as an execution time. The effectiveness of the proposed OFS-NN 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. 7 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 OFS-NNapproach on the basis of varied dataset size.

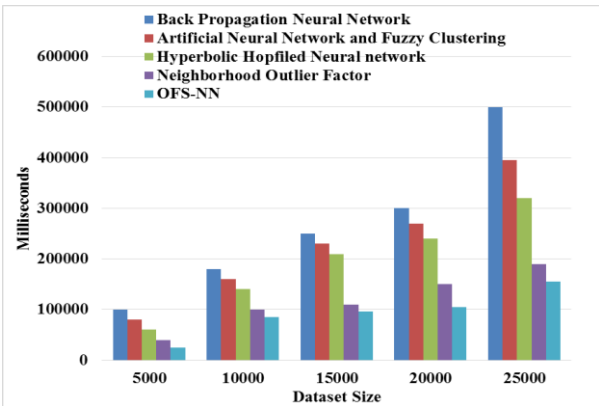


Fig. 7. Execution time analysis

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 25000 ms and 155000 ms respectively, which is proficient by 37.5 % and 18.42% respectively.

E. 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.

$$Accuracy\ (\%) = \frac{No\ of\ true\ outlier\ detections}{Total\ number\ of\ detections} \times 100 \quad (9)$$

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 OFS-NN approach owing to best feature selections made by applying ACO and sequence pattern clustering methodology.

The detection accuracy of the OFS-NN technique is compared against Memory Efficient increment Local Outlier Factor detection algorithm for more flexible version (MiLOF_F) [27] 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 OFS-NN efficiently reduces the computational complexity by using the optimal feature selection techniques and provides the accuracy up to 96.5 % for data dimension of 10 and 96 % of a data dimension of 50. Fig. 8 shows that the effective detection accuracy obtained and the deviation estimation in OFS-NN offers 2.65% and 3.125% better than the MiLOF-F respectively for minimum and maximum data dimensions.

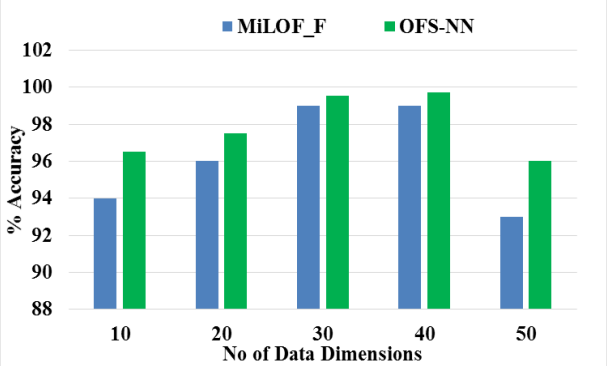


Fig. 8. Detection accuracy analysis

F. 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 OFS-NN and MiLOF_F approaches. As depicted in Fig.9 the proposed OFS-NN methodology necessitates only the least amount of memory than the existing MiLOF_F in a comparative manner.

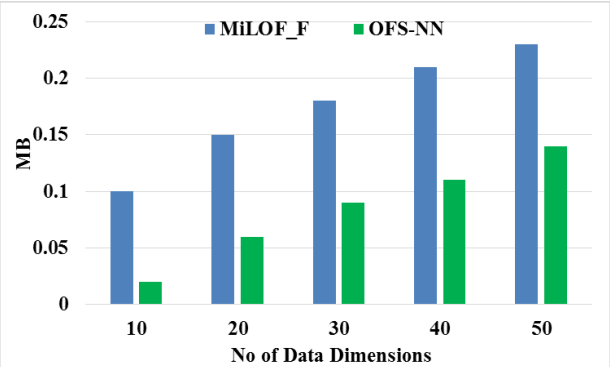


Fig. 9. Memory consumption analysis

The proficiency of the formulated OFS-NN 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 OFS-NN methodology exposes a memory consumption of 0.02 MB and 0.14 MB for data dimensions 10 and 50 respectively. Ultimately, OFS-NN exposes 80% and 39.13 % reduction than MILOF_F.

V. Conclusion

This paper proposed an Optimal Feature selection based Neural Network Model (OFS-NN) 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 ABC implying on a preprocessed data. The entire dataset is recognized in a SPARK architecture. After the preprocessing the subset separation take place. The separate subsets are send to feature selection take place by means of ABC algorithm. The expectation –maximization clustering algorithm is used for clustering process. The clustered datasets are classified as normal and outlier by using neural network classification. The efficacy of OFS-NN 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. OFS-NN evidently proves to be efficacious than other approaches in terms of mitigated execution time by 37.5% and 18.42% for the minimum and maximum dataset size accompanied with an enhanced detection accuracy of 2.65% and 3.125 for minimized and maximized data dimensions.

References

[1] D. Georgiadis, M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos, "Continuous outlier detection in data streams: An extensible framework and state-of-the-art algorithms," in *Proc. ACM SIGMOD International Conference on Management of Data*, 2013, pp. 1061-1064.

[2] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 2250-2267, 2014.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, p. 15, 2009.

[4] A. B. E. D. Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," *World Journal of Computer Application and Technology*, vol. 2, pp. 43-47, 2014.

[5] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 194-207, 2014.

[6] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "SVDD-based outlier detection on uncertaisn data," *Knowledge and Information Systems*, pp. 1-22, 2013.

[7] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for

unsupervised feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 2138-2150, 2014.

[8] P. Chalotra and M. Dutta, "An outlier detection method based on artificial bee colony fuzzy clustering," *International Journal of Advanced Research in Computer Science*, vol. 3, 2012.

[9] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 1369-1382, 2015.

[10] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, "Focused clustering and outlier detection in large attributed graphs," in *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1346-1355.

[11] M. Hubert, P. J. Rousseeuw, and P. Segaert, "Multivariate functional outlier detection," *Statistical Methods & Applications*, vol. 24, pp. 177-202, 2015.

[12] B. Liu, Y. Xiao, S. Y. Philip, Z. Hao, and L. Cao, "An efficient approach for outlier detection with imperfect data labels," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 1602-1616, 2014.

[13] D. Fleischhacker, H. Paulheim, V. Bryl, J. Völker, and C. Bizer, "Detecting errors in numerical linked data using cross-checked outlier detection," in *Proc. International Semantic Web Conference*, 2014, pp. 357-372.

[14] A. Zimek, M. Gaudet, R. J. Campello, and J. Sander, "Subsampling for efficient and effective unsupervised outlier detection ensembles," in *Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 428-436.

[15] X. H. Dang, I. Assent, R. T. Ng, A. Zimek, and E. Schubert, "Discriminative features for identifying and interpreting soutliers," in *Proc. IEEE 30th International Conference on Data Engineering*, 2014, pp. 88-99.

[16] Y. Liu, F. Nie, J. Wu, and L. Chen, "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion," *Neurocomputing*, vol. 105, pp. 12-18, 2013.

[17] W. C. Lin, S. W. Ke, and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-based Systems*, vol. 78, pp. 13-21, 2015.

[18] Q. Cai, H. He, and H. Man, "Spatial outlier detection based on iterative self-organizing learning model," *Neurocomputing*, vol. 117, pp. 161-172, 2013.

[19] D. G. Ferrari and L. N. De Castro, "Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods," *Information Sciences*, vol. 301, pp. 181-194, 2015.

[20] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, pp. 4324-4338, 2015.

[21] M. K. Siddiqui and S. Naahid, "Analysis of KDD CUP 99 dataset using clustering based data mining," *International Journal of Database Theory and Application*, vol. 6, pp. 23-34, 2013.

[22] Y. Yu, L. Cao, E. A. Rundensteiner, and Q. Wang, "Detecting moving object outliers in massive-scale trajectory streams," in *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 422-431.

[23] B. van Stein, M. van Leeuwen, and T. Bäck, "Local subspace-based outlier detection using global neighbourhoods," in *IEEE International Conference on Big Data*, 2016, pp. 1136-1142.

[24] D. Ienco, R. G. Pensa, and R. Meo, "A semisupervised approach to the detection and characterization of outliers in categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 1017-1029, 2017.

[25] W. N. S. Dr. William H. Wolberg and O. L. Mangasarian. (1993). *Breast Cancer Wisconsin (Diagnostic) Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[26] J. Jabez and B. Muthukumar, "Intrusion Detection System (IDS): Anomaly detection using outlier detection approach," *Procedia Computer Science*, vol. 48, pp. 338-346, 2015.

[27] G. Karthikeyan and P. Balasubramanie, "Robust feature selection model for outlier detection using fuzzy clustering and rule mining," *International Journal of Applied Engineering Research*, vol. 12, pp. 7019-7028, 2017.

G. Karthikeyan has completed his B.Tech (IT) in Vellore Institute of Technology (Madras University), Tamil Nadu and M.Tech (CSE) in Dr.M.G.R Research Educational and Research Institute, Tamil Nadu - India. He was worked in IBM India Pvt Ltd as a Data Specialist. He was worked as a Technical Lead in Wipro Technologies, Chennai and currently he is working as a Sr. System Analyst in UST Global Chennai Tamil Nadu. He is pursuing Ph. D in Anna University Chennai, Tamil Nadu under guidance of Dr. P. Balasubramanie Kongu Engineering College, Tamil Nadu. He has published 4 articles in International Conferences / Journals.

Dr. P. Balasubramanie is currently working as a Professor, Department of Computer science and Engineering, Kongu Engineering College, Perundurai, Tamilnadu India. He has completed his Ph.D degree in the year 1996. He is one of the recognized guide of Anna University, Chennai and Guided 25 PhD Scholars. Currently 10 scholars are pursuing their PhD degree under his guidance. He has Published 191 articles in reputed journals and many researchers cited/referred his work. He has published 11 books and three of the books published are used as text/reference book in many of the leading Universities. He is the recipient of the CTS Best Faculty award, P.K.Dass memorial Best faculty Award, Best Academic researcher award (ASDF),

Outstanding Scientist Award, Best researcher Award and Best Scientist Award.He has organized 21 SDPs/STTPs/Seminar/workshops for the benefit of faculty members and research scholars. He has successfully completed one AICTE RPS as a Principal investigator. He has organized several training programme for the benefit of students. He is actively involved in many software development works. His area of interest includes data mining, Networking and optimization Techniques.